Deep Learning-Based Automatic Analysis of Social Interactions from Wearable Data for Healthcare Applications

Ossama Sameer Alshabrawy

School of Computing
Newcastle University

This dissertation is submitted for the degree of
Doctor of Philosophy

Newcastle University December 2019

I would like to dedicate this thesis to my loving parents, my wife and my son …

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 45,000 words including appendices, bibliography, footnotes, tables and equations and has fewer than 150 figures.

Ossama Sameer Alshabrawy
December 2019

Acknowledgements

In the Name of Allah, the Most Beneficent, the Most Merciful,

First, and foremost, I am very grateful to my first supervisor Dr. Jaume Bacardit for the extensive support and encouragement through every single stage in my PhD journey. His outstanding and invaluable guidance, patience, inspiration, advice, and his extensive understanding of Machine Learning and remarkable enthusiasm for research has helped me to overcome many challenging situations in the research environment.

Special thanks goes to Prof. Thomas Ploetz for the extensive support in the first year he supervised me. I would also like to thank my second supervisor Prof. Patrick Olivier for giving me the chance to pursue a PhD at Newcastle University, supporting me to get Newton Fund and through the whole PhD program.

Special thanks goes to Bethany Little for helping with the statistical analysis and support. I wish to thank all the colleagues from ICOS research group. They helped me to see the brighter side of things and always gave me hope. I wish also to thank the School of Computing and Newcastle University. In addition, I very much appreciate Newton-Mosharafa Fund and the mission sector and cultural affairs, ministry of Higher Education in Egypt. Finally, I wish to thank Damietta University for giving me the chance to complete my studies in Newcastle University.

Abstract

Social interactions of people with Late Life Depression (LLD) could be an objective measure of social functioning due to the association between LLD and poor social functioning. The utilisation of wearable computing technologies is a relatively new approach within healthcare and well-being application sectors. Recently, the design and development of wearable technologies and systems for health and well-being monitoring have attracted attention both of the clinical and scientific communities. Mainly because the current clinical practice of – typically rather sporadic – clinical behaviour assessments are often administered in artificial settings. As a result, it does not provide a realistic impression of a patient's condition and thus does not lead to sufficient diagnosis and care. However, wearable behaviour monitors have the potential for continuous, objective assessment of behaviour and wider social interactions and thereby allowing for capturing naturalistic data without any constraints on the place of recording or any typical limitations of the lab-setting research. Such data from naturalistic ambient environments would facilitate automated transmission and analysis by having no constraints on the recordings, allowing for a more timely and accurate assessment of depressive symptoms. In response to this artificial setting issue, this thesis focuses on the analysis and assessment of the different aspects of social interactions in naturalistic environments using deep learning algorithms. That could lead to improvements in both diagnosis and treatment.

The advantages of using deep learning are that there is no need for hand-crafted features engineering and this leads to using the raw data with minimal pre-processing compared to classical machine learning approaches and also its scalability and ability to generalise. The main dataset used in this thesis is recorded by a wrist worn device designed at Newcastle University. This device has multiple sensors including microphone, tri-axial accelerometer, light sensor and proximity sensor. In this thesis, only microphone and tri-axial accelerometer are used for the social interaction analysis. The other sensors are not used since they need more calibration from the user which in this will be the elderly people with depression. Hence, it was not feasible in this scenario. Novel deep learning models are proposed to automatically analyse two aspects of social interactions (the verbal interactions/acoustic communications and physical activities/movement patterns). Verbal Interactions include

the total quantity of speech, who is talking to whom and when and how much engagement the wearer contributed in the conversations. The physical activity analysis includes activity recognition and the quantity of each activity and sleep patterns.

This thesis is composed of three main stages, two of them discuss the acoustic analysis and the third stage describes the movement pattern analysis. The acoustic analysis starts with speech detection in which each segment of the recording is categorised as speech or non-speech. This segment classification is achieved by a novel deep learning model that leverages bi-directional Long Short-Term Memory with gated activation units combined with Maxout Networks as well as a combination of two optimisers. After detecting speech segments from audio data, the next stage is detecting how much engagement the wearer has in any conversation throughout these speech events based on detecting the wearer of the device using a variant model of the previous one that combines the convolutional autoencoder with bi-directional Long Short-Term Memory. Following this, the system then detects the spoken parts of the main speaker/wearer and therefore detects the conversational turn-taking but only includes the turn taking between the wearer and other speakers and not every speaker in the conversation. This stage did not take into account the semantics of the speakers due to the ethical constraints of the main dataset (Depression dataset) and therefore it was not possible to listen to the data by any means or even have any information about the contents. So, it is a good idea to be considered for future work.

Stage 3 involves the physical activity analysis that is inferring the elementary physical activities and movement patterns. These elementary patterns include sedentary actions, walking, mixed activities, cycling, using vehicles as well as the sleep patterns. The predictive model used is based on Random Forests and Hidden Markov Models. In all stages the methods presented in this thesis have been compared to the state-of-the-art in processing audio, accelerometer data, respectively, to thoroughly assess their contribution. Following these stages is a thorough analysis of the interplay between acoustic interaction and physical movement patterns and the depression key clinical variables resulting to the outcomes of the previous stages. The main reason for not using deep learning in this stage unlike the previous stages is that the main dataset (Depression dataset) did not have any annotations for the speech or even the activity due to the ethical constraints as mentioned. Furthermore, the training dataset (Discussion dataset) did not have any annotations for the accelerometer data where the data is recorded freely and there is no camera attached to device to make it possible to be annotated afterwards.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background and Motivation

Late-life depression (LLD) is a common disorder associated with a decline or pervasive impairments in daily functioning [46]. Compared to depression in younger adults, LLD is associated with an increased burden of physical illness, more impaired functioning, more severe neuropsychological impairment, particularly in executive and psychomotor functioning and a poorer clinical outcome [46, 165]. The clinical practice assessments of depression often rely on qualitative measures as a result of analysing face-to-face interviews, meeting scenarios, diaries, questionnaires and surveys. These artificial settings are tended to be biased and do not provide a realistic impression of the patient and affected by the patient's mood. Thus, These clinical assessments don't lead to sufficient diagnosis and care. Previous research has demonstrated the utility of wearable technology which have advanced rapidly enabling, unobtrusive objective long-term monitoring, the provision of feedback and inference of behavioural changes by individuals in the home and community settings [125]. O'Brien et al and Prince et al. have used wearable behaviour monitors (e.g. actigraphs) to objectively measure physical activity in participants with LLD, with these methods producing more accurate measures than the current clinical practice [121, 131]. Consequently, it has been suggested that wearable technology could be useful in more objectively quantifying social activity in participants with LLD and, specifically, that wearable devices could detect speech activity and physical activity that an individual is exposed to and engages in, as an ecologically valid measure of social interaction [74].

Data provided by these wearables allows remote monitoring, which provides more realistic data as behaviour is captured in naturalistic-setting environments rather than in often artificially clinical settings [176]. Such sensors are capable of measuring a multitude of physiological parameters (e.g., heart rate, blood pressure, body and skin temperature,

oxygen saturation, respiration rate, electroencephalogram, speech, and body movements [125]. Existing applications of wearable sensing systems and behavioural data monitoring includes healthcare assessment analysis. These applications to healthcare and well-being can include - but are not limited to - acoustic analysis, gait analysis, fall detection and sleep monitoring [15].

The social interactions analysis has two main aspects: verbal interactions/acoustics communications and physical activities/movement patters. The scientific literature [27, 11] of human social interactions analysis systems includes the design of a system that utilises the information provided by sensors such as tri-axial accelerometer (providing movement patterns) or microphone (providing speech). Such a system combines mobile and direct sensing with pattern recognition along with machine learning/deep learning techniques (such as logistic regression, support vector machines, random forests and deep neural networks) for modelling and inferring physical activities or acoustic social communication [11, 140]. Existing automated speech analysis methods in conversations modelling generally take into consideration very limited situations that poorly reflect natural emotions such as in meeting room scenarios [7]. Automated analysis of specific acoustic features of speech can distinguish participants with depression from controls with accuracy levels of 75-80% [183], with the former showing shortened voice onset time, decreased second formant transition, and increased spirantisation [99, 145].

As in countless other applications, deep learning has shown promise in the processing of data generated by wearable devices. Deep learning (DL) is a representation learning class of machine learning algorithms that is robust, generalizable and scalable. Robust means that there is no need to hand crafted features engineering. The features are extracted/learned from the raw data through multiple levels. Each level represents abstract set of features and each level corresponds to a layer in the artificial neural network to produce the output. This leads to minimal pre-processing of the raw data. DL is also robust to the natural variations in the data that automatically learned. Furthermore, deep learning has the capability to deal with big data and are flexible and big to store enough knowledge from the data and hence are capable to generalise. Also, DL architectures can be used for many different applications through transfer learning. Moreover, deep learning has powerful priors to solve the curse of dimensionality problem. For example, in speech analysis, unsupervised deep learning models could be utilised as a pre-training process to extract features using deep belief networks [188]. Convolutional neural networks are used for speech recognition such as in [166]. Recurrent neural networks and long short-term memory (LSTM) and bidirectional LSTM have been thoroughly applied in speech analysis [76, 85, 97].

The other aspect of social interactions is the physical activity analysis also known as human activity recognition (HAR). HAR uses the wearable devices to monitor and assess/analyse the everyday life activities of the users using mainly accelerometer, gyroscope and compass sensors. These sensors collect raw data from body, wrist, ankle, etc. for a variety of purposes, for instance, classifying the individual's behaviour into four activities: ambulation, cycling, sedentary and other [105]. This thesis focuses on predicting cycling, sedentary, light and moderate tasks, vehicle usage and sleep patterns.

Automated analysis of social interactions especially for wearable data collected in naturalistic settings environment still remains very challenging. On the contrary to controlled settings environment (e.g. clinical/lab settings), naturalistic environment presumes no constraints on where the data is recorded or how the social interaction happen. This research uses a wrist-worn device that comprise a microphone and a 3D accelerometer (details of the device is discussed in Chapter 3) to record speech and movement data for people with depression. This data will be objectively analysed to quantify the social-functioning and extract digital markers for depression.

1.2 Overview of the Problem

Finding robust objective measures of the social functioning is a demanding task to help the psychiatrist to make better diagnosis and treatment. The utilisation of wearable wrist-worn devices provides the opportunity to record audio and movement data in naturalistic settings, potentially leading to the generation of better measures of social interactions. These settings cause more challenging types such as audible speech in the background (from people who are not part of the conversation or from devices that generate speech such as TVs and radios) of noise especially in the audio data as well as the data is privacy-sensitive and hence not annotated. Due to the privacy issues, it will not be possible to listen to the data by any means. In response to these issues, this thesis will address these challenges.

1.3 Aim and Research Questions

The central hypothesis in this dissertation is that social interactions of LLD can be measured or quantified using wearable sensing platforms and machine learning models. Social interactions comprise verbal interactions or other acoustic communications, and general activities or movement patterns. The underlying assumption is that social interactions are an important aspect for the characterisation of an individual's social functioning [156], and are very relevant for assessing people with LLD in a quantitative manner. This important assessment

of social functioning aspects help the psychiatrists improve both the diagnosis and treatment. Social interactions will be objectively measured by machine learning algorithms base on a data captured by a wearable device that is intended to be used without any constraints and easy to use and could be worn everywhere. Consequently, the main aim of this dissertation is to design, develop, evaluate algorithms for the objective assessment of social functioning well-being with no constraints on the recordings of the sensed data. Then the social functioning is used as a digital markers for depression and therefore discriminate between healthy controls and depressed cohort. The research questions are listed below:

1. How can the total amount of verbal/acoustic interactions be robustly assessed in naturalistic settings? The audio recordings produced by the wrist-worn device include speech and non-speech patterns. The speech belongs to the wearers and other people interacting with them. In order to objectively measure the total amount of verbal interactions, it is required to detect the speech precisely excluding all the non-speech patterns. The non-speech patterns are the background noise, audible speech within background.

2. How can we objectively analyse the quantity of verbal/acoustic interactions for the wearer of the wrist-worn monitoring device? After the total speech is detected, the substantial point now is to infer who is talking to whom and when, how much and how often does the wearer engage in any conversation. This requires the development of robust methods for automatically detecting the wearers (main speakers) from the conversation and infer the quantity of speech the device wearers produced. This is followed by a statistical analytics that shows that speech measures could be used as digital markers for depression especially for LLD.

3. How can we robustly assess physical activities/movement patterns of the wearer? Social functioning well-being relates not only to the interactions through spoken language but also to a physically active lifestyle. As a second dimension of the automated assessment of social functioning, robust analysis methods that infer elementary physical activities and movement patterns from tri-axial accelerometer data will be developed. Such elementary activities include ambulatory modes of transportation (bicycle, bus, car, etc.), ambulatory activities such as walking, light and moderate tasks, sedentary actions as well as the sleep patterns.

4. How do the spoken interactions, physical movement patterns and the key clinical variables for depression correlate with each other? Psycho-social functioning is characterised by a healthy degree of both verbal interactions with other people and movement

patterns that indicate an active lifestyle. So, the objective here is investigating the correlation between the speech measures and the key clinical variables for depression, the correlation between the physical activities and also the same key variables and finally the correlation between the speech measures and the physical activities.

1.4 Main Contributions

The main contribution of this dissertation is the introduction of a novel approach to the automated analysis of social interactions of people with LLD in order to provide the clinician with additional information to make better diagnosis and treatment. This approach is a result of the critical analysis of the state-of-the-art speech detection, speaker segmentation and human activity recognition methods. This is directly addressing two main issues: (1) the data recorded for people with depression is privacy-sensitive (2) the data is recorded in naturalistic settings (i.e. there is no constraints on where the wrist-worn device should be used or any other restrictions on the recordings). That leads to challenges of having unannotated dataset and challenging types of noises generated by recording in naturalistic environments. The proposed solution comprises three main elements, each described in a separate chapter:

- A novel robust deep learning architecture for speech activity detection that utilises gated bi-directional LSTM and maxout networks with a combined optimiser. This architecture has shown to have superior speech detection performance, especially in processing audio from naturalistic settings. In addition, the model is able to discriminate between the speech of the wearer and people interacting with them and the audible speech within background. From the model predictions, the speech measure is quantified as a percentage of speech averaged over the week (the length of wear time for each participant) and used as a digital marker for depression and shows significant difference between the healthy controls and depressed cohort.

- A novel deep learning model for detecting the wearer's speech from the whole predicted speech utilising convolutional autoencoder and bi-directional LSTM and a combined optimiser. This model is capable of differentiating between the wearers and other speakers interacting with them. This is basically different from speaker segmentation where only the wearer is the main focus in order to infer how much they engage with other speakers during conversation. This also used as another significant marker for depression.

- Extracting meaningful high level activity features from a pre-trained model to obtain new digital markers from the physical activities and analyse the correlations with the

key clinical variables used by the psychiatrists to make the diagnosis and develop a machine learning model for discriminating between the healthy controls and depressed patients with higher performance and thorough evaluation with nested cross validation.

1.5 Research outputs

1.5.1 Publications

Bethany Little, Ossama S. Alshabrawy*, Daniel Stow, Nicol Ferrier, Thomas Ploetz, Jaume Bacardit, Patrick Olivier, Peter Gallagher, John O'Brien, "Deep learning-based automated speech detection as a marker of social functioning in late-life depression", Psychological Medicine (IF: 5.641). (*equal first authorship)* This paper represents the application part of Chapter 3 and Chapter 4.

1.5.2 Publications under review

Ossama S. Alshabrawy, Bethany Little, John O'Brien, Peter Gallagher, Patrick Olivier, Jaume Bacardit, "NatSpeech: Deep Learning based Speech Activity Detection in Naturalistic Environments", IEEE Transactions on Pattern Analysis and Machine Intelligence (IF: 17.730, under review). This paper represents the main core of Chapter 3.

1.5.3 Publications in preparation

Ossama S. Alshabrawy, Peter Gallagher, John O'Brien, Jaume Bacardit, "Automated Assessment and Digital Markers of Physical Activities for People with Late-Life Depression" (paper in preparation). This paper will cover the work of chapter 5.

1.6 Structure of the Thesis

The dissertation is organised as follows. Chapter 2 explains the wearable computing technology and its applications in healthcare and well-being. Subsequently, an overview of the deep learning models which have been widely used in social interactions which includes speech analysis and human activity recognition. Then, the speech analysis are discussed in details including the data. Finally, human activity recognition state-of-art is presented.

Chapter 3 describes a robust speech measure for the verbal interactions/acoustic communications problem. A novel deep learning architecture with a thorough comparisons of the proposed model with various state-of-art methods. This measure is used to significantly

discriminate between the healthy control subjects and depressed patients. This measure is later used in constructing the speech events in the conversation in Chapter 4. Also, the correlation with the key clinical variables of depression diagnosis is investigated.

Chapter 4 investigate how much and how often the wearer of the wrist-worn device is engaged in any conversation. A novel deep learning approach is discussed for detecting the wearer from the predicted speech segments. The amount of speech produced by the wearer is also used as a digital marker for depression. Moreover, A thorough validation and evaluation to the model is presented followed by statistical analysis for the digital marker and its correlation with the key clinical measures.

Chapter 5 describes the method for predicting physical activities and a thorough analysis for some activity markers from the predictions. Furthermore, the correlation with the clinical measures is also presented. Moreover, the correlation with the speech markers presented in Chapter 3 and 4 is discussed. Finally, new high level features extracted from the physical activity predictions are extracted and used to discriminate between the healthy controls and the depressed cohort is also described.

Chapter 2

Background and Related Work

This chapter introduces the concepts of the wearable computing and sensor data with application to social interactions and healthcare. This is followed by an introduction to machine learning followed by detailed background about deep learning. The next part of the chapter represents the application of machine learning/deep learning to human activity recognition. The application of human activity recognition includes speech and physical activity as the main aspects of social interactions. The last section discusses the classification performance measure that is used through the thesis.

2.1 Wearable Computing for Social Interactions & Healthcare

2.1.1 Wearable Computing

The continuously increasing popularity of wearable computing technologies such as smartwatches and fitness/well-being bands suggest that wearable healthcare and well-being devices are a promising research [11]. Contemporary healthcare applications are evolving and wearables are becoming progressively smaller and much more energy efficient, making them relevant for continuously objective sensing and giving feedback. The inherent portability of wearable health and well-being monitoring enables unobtrusive and accurate data collection, which can lead to long-term healthcare monitoring [125]. Additionally, wearable sensors applications can harness the power to drive behavioural changes [176].

Data provided by wearable devices allows remote monitoring, which is not only more comfortable for the patients and cuts costs for healthcare systems [176], but also provides more realistic data as behaviour is captured in naturalistic-setting environments rather than in

often artificially clinical settings. Longitudinal ambient monitoring systems of behavioural data is an important common objective application and has motivated a variety of different developments of sensing systems, ranging from wearable systems (based on inertial and other sensors) to ambient systems (which make use of video cameras, motion sensors, etc.) [83]. Smartphone-based wearable sensing platforms have been considered to be effective means for the analysis in the computational social sciences [174, 107].

2.1.2 Applications to Healthcare

Existing applications of wearable sensing systems and behavioural data monitoring includes healthcare assessment analysis [15] and activity-related analysis. Health monitoring may comprise of various types of miniature sensors. Such sensors are capable of measuring a multitude of physiological parameters (e.g., heart rate, blood pressure, body and skin temperature, oxygen saturation, respiration rate, electroencephalogram), and body movements [125]. These applications to healthcare and well-being can include - but are not limited to - gait analysis, fall detection and sleep monitoring. Monitoring sleep is a pivotal biomarker of the human general health. This quality analysis comprises the deepness of sleep, the duration and regularity. The ground truth for sleep analysis is often acquired by self-reports or camera worn within the process of recording and then the video can be used to manually annotate the streaming data. Laerhoven et al. [93] rely on light sensors and tilt switch sensor to predict the posture of sleep and motion duration within the night. The findings they come up with are that the nine tilt switch sensors have a strong correlation with the posture of the body and the accuracy of their system is 80%.

Unlike using sensors that are attached to specific part of the body, smart-phones are also utilised for the same purpose. Min et. al. [111] proposed a model that uses microphone, proximity sensor, 3D accelerometer, and other sensors from the smart-phone in order to classify the sleep and wake times. Then based on the prediction of sleep and wake times, they interpret the good and bad sleep. The authors in [22] build another model also on smart-phone but the difference between them and [111] is that they consider the light condition and duration of the phone usage and the silence factor based on the built-in microphone.

2.1.3 Applications to Social Interactions

According to the literature, human social interaction analysis includes the design of a system that utilises the information provided by sensors which determines whether they are wrist-worn, the quantity and the type of sensor such as tri-axial accelerometer or microphone [11]. Such a system combines mobile, direct sensing with pattern recognition and machine learning

techniques for modelling and inferring physical activities or acoustic social communication [11]. The integration of human social interaction analysis with wearable computing aims at providing more flexible and naturalistic computing technology [171] and therefore allows the automatic social interaction analysis. Essentially, existing reported speech analysis research in conversations modelling generally take into consideration very limited situations that poorly reflect natural emotion such as in meeting room scenarios [7]. For instance, the captured real emotion datasets [120, 39] are restricted to a limited number of observations and are also recorded in relatively non-naturalistic settings that do not represent daily life.

The increase in life expectancy nowadays in addition to ageing changes which leads to physical or cognitive decline has a substantial influence on the everyday life for people [121]. That in essence will in turn lead to injuries, mental issues or shrinkage of the levels of physical activities [121]. The activity-related applications use the wearable devices to monitor and assess/analyse the everyday life activities of the users. The wearable sensors used for that are mainly accelerometer, gyroscope and compass. The accelerometer measures the acceleration in the three orthogonal axes. Gyroscope measure the angular velocity in the Cartesian coordinates whereas compass is measuring the orientation in geodetic coordination system. The activity analysis includes step counting, exercise statistics, gesture recognition, indoor navigation, rehabilitation after neurological disorder such as a stroke. The rehabilitation is considered as a long process towards human motion mobility, tracking, regaining function, and independence based on the movement data (accelerometer sensor data) with data annotations [154, 49]. This process can be more challenging if it is without data annotations [147], or home-based rehabilitation exploiting [119]. On the other hand, there are more applications such as detecting whether the user quit smoking using the respiratory rate based on the movement data of the hand and chest while sleeping [81]. Off the large-scale data activity projects, there are UK Biobank and NHANES that are utilising wrist-worn accelerometer-based activity monitors that collect raw data for the purpose of processing wrist and ankle raw data and classify the individual's behaviour into four activities: ambulation, cycling, sedentary and other [105].

Deep learning models have been widely used in social interaction analysis which include the verbal interactions/speech analysis and physical activity analysis as the two main aspects of social interactions. This will be presented in the next sections. The next section introduces classical machine learning followed by deep learning models. Then this is followed by speech analysis and activity recognition that will be presented in detail in section 2.4.

2.2 Introduction to Machine Learning

In this section, introduction to classical machine learning is introduced. This introduction includes supervised and unsupervised leaning concepts followed by the classical supervised methods used in this thesis. These methods are support vector machines and random forests.

2.2.1 Supervised vs Unsupervised Learning

The learning tasks in machine learning can be classified into three categories: supervised, unsupervised and reinforcement learning [16]. Supervised learning is the machine learning task that learns a function that maps the samples of data to its desired output. In other words, the samples of the data has a ground truth (i.e. there is prior knowledge for the data samples) [16]. The supervised learning is typically used in classification or regression tasks. A supervised learning model analyses the training samples from the data and produces an inferred mapping to label the output for new samples that are unseen by the model (i.e. the model generalise to other data points). The most common algorithms for supervised learning includes linear regression, logistic regression, support vector machines, decision trees, random forests and deep neural networks/deep learning. In this thesis, only support vector machines and random forests are introduced and the deep learning algorithms will be introduced with details in Section 2.3.

On the other hand, unsupervised learning is a machine learning task that the data samples do not have a ground truth [68]. In contrast to supervised learning that uses the human-labelled data, unsupervised learning allows for modelling the probability density of the data samples. This means that the main aim is to learn the inherent structure of the data without using explicitly-provided labels [16]. The most tasks for unsupervised learning are clustering, representation learning and density estimation. The common algorithms for unsupervised learning includes k-means clustering, principal component analysis and autoencoders.

2.2.2 Support Vector Machines

Support vector machine (SVM) is a supervised learning model that can be used for classification, regression analysis or even clustering [16]. However, it has been widely used in classification tasks. The idea of SVM is that is to obtain a hyperplane which discriminate the data points distinctly. This hyperplane is n-dimensional space where the number of dimensions n is specified based on the number of features in the data. Separating two classes by a hyperplane has many possible choices so, the main aim of SVM is to find the hyperplane that has the maximum margin [31]. The maximum margin means the maximum distance between

data points of both classes. Generally speaking, the hyperplanes are decision boundaries used to classify the data points to decide whether the data point falls on either side of the hyperplane. For instance, if the number of features is 2, then the hyperplane is just a line and in case of 3 features that requires a two-dimensional plane. The question arises now is what the support vector are. Support vectors are the closest data points to the hyperplane and affect its position and orientation [31]. These support vectors are used to maximise the margins of the hyperplane. The large margin is different from logistic regression where the decision is made based on 0.5 threshold (if the output > 0.5, then it will be assigned to one class and otherwise it will be assigned to other class). However, in SVM the output of the linear function used with SVM varies between [-1,1] [16]. The loss function that is used to maximise the margin (i.e. soft margin) is hinge loss which is defined in Equation 2.1.

$$\begin{cases} 0 & y * f(x) \geq 1 \\ 1 - y * f(x) & otherwise \end{cases} \tag{2.1}$$

Any hyperplane can be written as a set of data points $\overrightarrow{x}$ that satisfies $f(x) = \overrightarrow{w}.\overrightarrow{x} - b = 0$, where $\overrightarrow{w}$ is the normal vector to the hyperplane. This illustrated in Fig. 5.1. So, the main objective is to minimise the function (Eq. 2.2). This equation represents the quadratic optimisation problem (also know as primal problem) [178].

$$arg\ min_{w,b,\xi_i} \frac{||w||}{2} + C\xi_i \tag{2.2}$$

where, $\xi_i = max(0, 1 - y_i f(x_i))$. If the data is not linearly separable, then two possible solutions. One solution is projecting the raw data into linearly separable feature space. This can be achieved by deep learning for feature extraction or by feature engineering. Thereafter, the linear SVM will be applied. The second solution is to use non-linear SVM (either the dual problem or the kernel trick) [143]. Representer theorem [143] states that w can be rewritten as a linear combination of the training data points as:

$$w = \sum_{j=1}^{N} \alpha_j y_j x_j \tag{2.3}$$

where, N is the number of features. The kernel trick is relies on kernel function $K(x_j, x_i) = \Phi(x_j)^T \Phi(x_i)$. The common examples for the kernel functions are linear kernels, polynomials kernels and radial basis function (RBF). RBF is the kernel used in this thesis through the experiments in Chapter 3.

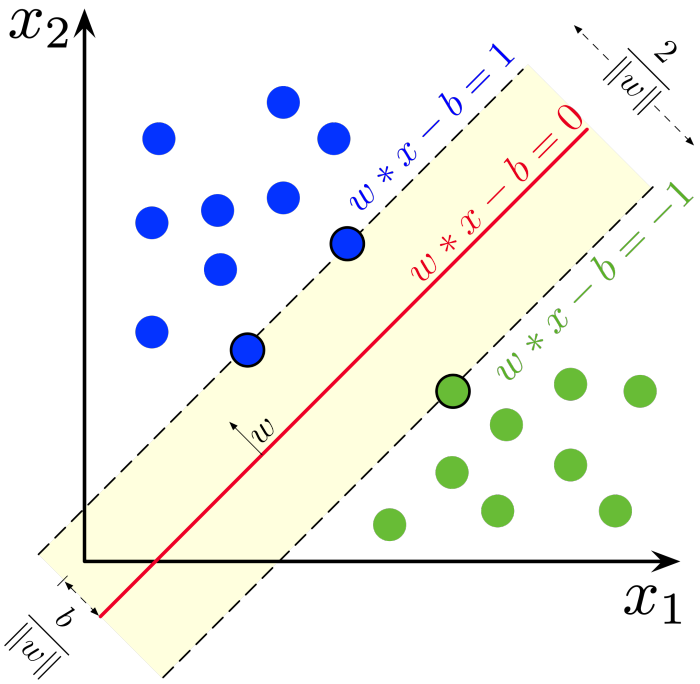

Fig. 2.1 Maximum-margin hyperplane illustration for SVM. SVM is trained on two classes. The support vectors are the samples on the margin.

2.2.3 Random Forests

A random forest (RF) is an ensemble algorithm and is comprised of multitude of decision trees. Random forests is commonly used algorithm for classification and regression tasks. Fig. 2.2 illustrates how random forest algorithm works for classification. Each tree is constructed with randomly selected number of examples with replacement. Subsequently, a set of randomly selected features but without replacement is used to split each node of the tree. Classification and regression trees (CART) are used in random forest. After that CART is applied to the samples and features to make predictions. The predictions from each tree is one of the activity classes. The ground truth labels for the classes are converted to one-hot encoding vectors. Each of the trees is giving a vote for only one class. The CART has the bagging property where the votes from the trees have a high variance and therefore these votes are combined. In order to predict the final class, the prediction is the class which has majority of votes. Another way for making the prediction is to normalise the votes by dividing the votes by the total number of trees [13]. This definitely will create a probability rather than producing the crisp class label.

Random forest has a high capability of classifying the instances in the dataset that does not have temporal dependence property. So, random forest is not designed to handle the temporal variation in the data (time-series specifically). In order to handle the temporal

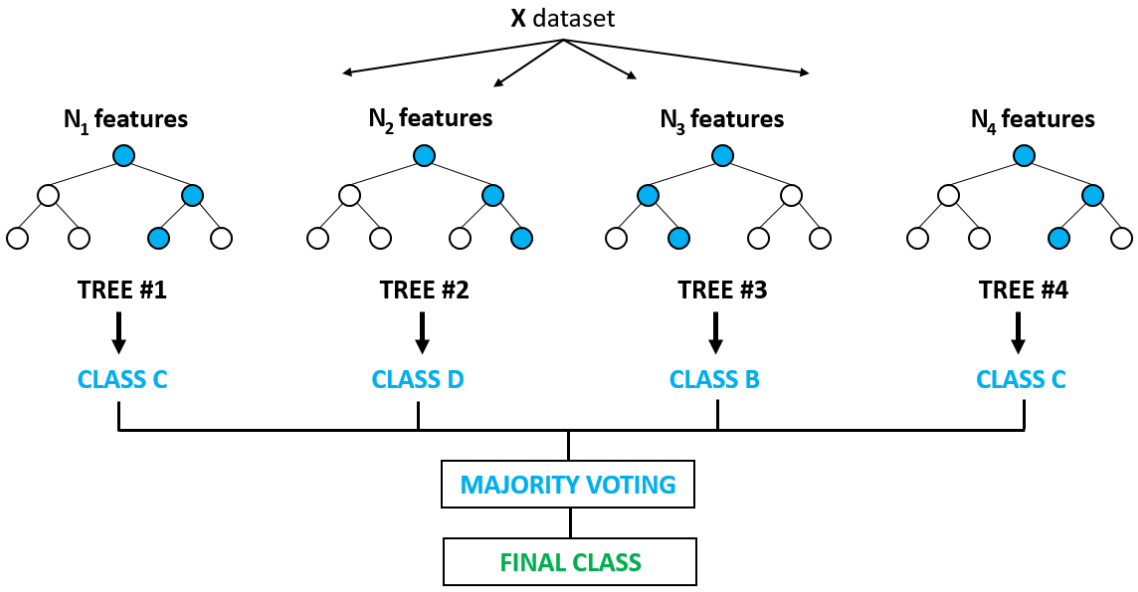

Fig. 2.2 Random forest algorithm illustration.

dependence, Gaussian mixture models or hidden Markov models are good options for that purpose. An application of using RF with hidden Markov models for activity recognition will be introduced in 2.4.7 and will be used in Chapter 5.

2.3 Deep Learning (DL)

Deep learning is part of machine learning (ML) and is a broad research field that is fundamentally based on artificial neural networks. Similar to ML, the learning process can be supervised, unsupervised or semi-unsupervised [54]. Deep learning has tremendous applications that includes but not limited to computer vision, bioinformatics, drug design, medical image analysis, material inspection and board game programs, speech recognition, audio recognition, natural language processing, audio recognition, social network filtering, machine translation, bioinformatics, drug design, medical image analysis, material inspection and board game programs [28, 90, 56]. In the last decade, the results of the deep learning methods were superior to human experts which put deep learning as a promising research field to solve complicated problems. In this thesis, my main focus is on deep learning models and their performance with the raw data without any feature learning. This section will present an overview of multilayer perceptron, optimisation algorithms for DL, recurrent neural networks, convolutional neural networks and autoencoders.

2.3.1 Multi-Layer Perceptron

Artificial Neural Network (ANN) were developed as mathematical models that simulate the capabilities of biological neural networks [108]. ANN consists processing units or nodes that are connected to each other through the weights. The nodes are equivalent to the neurons in the biological neural network and the weight connections are equivalent to synapses between the neurons. Stimulating the neurons with some input makes the network spread out activation along the connections. ANN has two types of connections: cyclic (feedback) connections and acyclic connections. ANN without feedback connections is called deep feed-forward network, or feed-forward neural network, or multilayer perceptron (MLP) [16]. MLP is illustrated in Fig. 2.3. The nodes in presented in layers feeding forward from a layer to the next layer starting from the input layer which contain the input patterns. The activations of the nodes are propagated through the hidden layer until it reaches the last layer (output layer) and this is called forward pass. The hidden units have activation function that maps the summed activation of all the input to that unit. These activation function are no-linear functions such as piecewise linear, hyperbolic tangent, sine, logistic, rectified linear unit, etc. The weighted sum of the input units is stand for the network input, then the activation function f is then applied to that weighted sum.

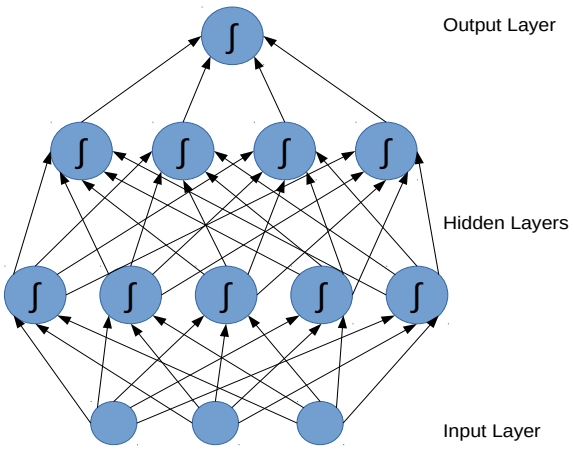

Fig. 2.3 MLP structure

Suppose we have input units from 1 to I and denote the weight from unit i to unit k as w_{ik}, then the weighted sum will be,

$$s = \sum_{i=1}^{I} w_{ik} x_i \tag{2.4}$$

then the activation function is applied as follows,

$$activation = f(s) \tag{2.5}$$

Once the activation of the first hidden layer is calculated, then the process is repeated until the last layer which is called output layer. The output vector relies on the number of units in the activation output layer and the choice of the activation function and both depends on the task. So, for a binary classification, the common activation function is the sigmoid (logistic) function and since the logistic function range is the open interval $(0,1)$, then the activation can be interpreted as the probability that input vector belongs to the first class or the second class [54]. For multi-class classification problem the softmax is chosen and it produces values that add up to 1.

MLP is trained using the backpropagation that fine-tune the parameters by approximating a cost/loss function. The choice of the cost function is substantial in the design of the ANN. The cost functions used in ANN models is quite similar to the other parametric models. In most cases, these models defines a distribution and then use the principal of the maximum likelihood and this what is used for modern neural networks. Usually the cost functions is called loss function where it tries to calculate the classification error and definitely aims at minimising that error. Since the *log* function is monotonic increasing function, this will be equivalent to the negative log-likelihood. The negative log-likelihood is also known as cross-entropy C [16] and it is given by:

$$C = -\sum_{s \in S} p \; log \; y + (1-p) \; log \; (1-y) \tag{2.6}$$

where, y is the actual value and p is the predicted value for each of the states $s \in S$. The states here represent the examples from the training data.

2.3.2 Backpropagation in Deep Learning and Network Training

The training of ANN can be performed by backpropagation which aims at minimising the differentiable cost function to train its parameters. This can achieved by gradient descent optimisation. The gradient descent starts with finding the derivatives of the cost/objective function with respect to the weights of the ANN and then the weights should be updated in the direction of the negative slope. Optimisation of ANN is different from the conventional optimisation where in machine learning we care about some performance measure. So, we minimise the cost function hoping that the performance measure will improve [54]. There

are a variety of solvers for the backpropagation process but only th most common algorithms are introduced here.

Stochastic Gradient Descent

Stochastic gradient descent (SGD) is an iterative algorithm to optimise a differentiable cost function. Basically, SGD is the stochastic approximation of the well known gradient descent algorithm. Unlike gradient descent which compute the gradient for all the examples in the entire dataset, SGD computes the gradients of randomly selected subsets of the data [12]. So, this is the reason it is called stochastic. The idea of SGD is to calculate the average gradient on n of examples considered as a minibatch (see equation 2.7. These examples are identically independent distributed from the data. This leads to an unbiased estimate of the gradient. The most critical parameter of SGD is choosing the learning rate η and commonly set as a fixed value. However, practically speaking the learning rate should be gradually decreased over the training set and through epochs. The term epoch in DL stands for a full pass over the dataset. That is, training all the minibatches in the datasets is considered to be one epoch and this will be repeated until the stopping criterion has met. The stopping criterion can be the number of epochs or early stopping approach. Early stopping means that to stop once the validation set error begins to rise and return to the validation parameter setting with the lowest validation error. This is can be achieved by saving a copy of the model parameters at each epoch [54]. The update rule for SGD is given by Eqs. 2.7 and 2.8.

$$\widehat{grad} \leftarrow \frac{1}{n} \nabla_{\omega} \sum_{i=1}^{n} L\left(f(x^{(i)}; \omega),\, y^{(i)}\right) \tag{2.7}$$

$$\omega \leftarrow \omega - \eta\, \widehat{grad} \tag{2.8}$$

where, ∇_{ω} is the gradient with respect to the network parameters, L is the loss function, x is the input, f is the activation function, ω are the parameters/weights of the neural network model, n is the batch size (number of examples in a mini-batch) and η is the learning rate. The first equation represents the calculation of the gradient with respect to the network parameters in order to be used when updating the parameters. There are various ways to choose the learning rate such as in Smith et al. [151] where this method assumes to have learning rates values within a reasonable boundary and then the learning rate will cyclically change within that range lets the learning rate cyclically vary between reasonable boundary values. Smith et al. argument is based on [Dauphin et al.] that the minimisation issue of a loss function is not only because of the local minima but also because of the saddle points. The saddle points will make the learning process very slow if the gradients are very small.

However, when increasing the learning rate, this leads to more traversal values of the saddle point which is also an issue. So, when using a range of learning rates and then repeating them cyclically, the loss function optimisation will lead to better performance. The standard way to choose the boundary range is to run the experiments with values of the learning rate such as 10^{-1}, 10^{-3}, 10^{-6}. According to the experiments done through this thesis, I applied the standard way rather than the cyclical one since this method needs intensive computation and also the learning is not the only parameter that affect the training. This standard way is to run with high, medium and low learning rate (10^{-1}, 10^{-3}, 10^{-6}) in order to obtain an overview of which rate is suitable for the training data and then decide the most appropriate one without using fixed learning rate but adaptive learning rate which will be discussed in the following subsections.

SGD with Nestrov momentum

The momentum algorithm proposed by Polyak 1964 [130] considers calculating the past gradients and then accumulates an exponentially decaying moving average of these gradients. Nestrov momentum algorithm [161] is a variant of the momentum that exploits the Nestrov's accelerated gradient method [118]. The update rule in this case is presented by:

$$\vartheta \leftarrow \beta\ \vartheta - \eta \nabla_{\omega} \left[\frac{1}{n} \sum_{i=1}^{n} L\left(f(x^{(i)}; \omega + \beta\ \vartheta),\ y^{(i)} \right) \right] \tag{2.9}$$

$$\omega \leftarrow \omega + \vartheta \tag{2.10}$$

where, ϑ is the velocity of convergence through the parameter space, β is a hyperparameter ranges from 0 to 1. From the first equation, ϑ is the the negative gradient exponential decaying average. This determines how fast is the exponential decay of the previous gradients. The weights are often initialised by random values from Gaussian or uniform distributions. The scale of these distributions has a great influence on the way the network is going to generalise. Choosing large weights will usually lead to get rid of the redundant units. This also helps not to lose important information form the raw data within the training [54].

AdaGrad

As discussed earlier in this chapter that the learning rate has a great influence on the model performance. Setting the learning rate is a challenging issue. Momentum method has tackled this issue but unfortunately it relies on the velocity which itself relies on β and thus the problem persists. In response to this issue, adaptive learning rate can be a solution. The

idea of adaptive learning rate is that instead of setting a fixed value for the whole model, adjust a learning rate for each parameter and within the training these learning rates will be adapted automatically. AdaGrad is one of these methods and basically it is a convex optimisation method. According to [41], AdaGrad is adapting the learning rates for all the model parameters by rescaling them inversely proportional to the square root of the sum of all the historical squared values of the gradient. The idea is that the learning rate is going to decrease but with the larger partial derivative of of the loss function, the decrease will be rapidly and the decrease in learning rates will be very small if the derivatives are small. The update rules for AdaGrad are given by the following equations:

The gradient is calculated like in 2.7, then the squared gradient (look at equation 2.11) will be accumulated:

$$sq\_grad \leftarrow sq\_grad + \widehat{grad} \odot \widehat{grad} \tag{2.11}$$

$$\triangle \omega \leftarrow -\frac{\eta}{\delta + \sqrt{sq\_grad}} \tag{2.12}$$

where δ is a small constant for numerical stability. Note that the operations in the previous equations are element-wise.

$$\omega \leftarrow \omega + \triangle \omega \tag{2.13}$$

RMSProp

Unlike AdaGrad, RMSProp [66] is working in a non-convex setting. The gradient accumulation in RMSProp is an exponentially decayed weighted moving average. This helps to exclude the extreme past history. Hence, it will then converge quicker [54]. The update rules is given by the following equations:

The gradient is calculated by 2.7, then the accumulated squared gradient is calculated differently from AdaGrad where RMSProp is using a decay rate ρ as follows:

$$sq\_grad \leftarrow \rho(sq\_grad) + (1-\rho)\widehat{grad} \odot \widehat{grad} \tag{2.14}$$

$$\triangle \omega \leftarrow -\frac{\eta}{\delta + \sqrt{sq\_grad}} \tag{2.15}$$

$$\omega \leftarrow \omega + \triangle \omega \tag{2.16}$$

Adam

Adam [87] method name is derived from Adaptive Moments estimation. Adam can be looked at as a combination of RMSprop and Stochastic Gradient Descent with momentum. However, the momentum in Adam represents an estimation of the first-order moment of the gradient with exponential weighting. Likewise, first and second order moments calculation, Adam incorporates bias correction which accounts for the initialisation of the moments at the origin. The update rules is given by the following equations:

The gradient is calculated by 2.7, then the first-order and second order momentum are calculated by the following two equations:

$$first\_moment \leftarrow \rho_1 \, first\_moment + (1-\rho_1)\, \widehat{grad} \tag{2.17}$$

$$second\_moment \leftarrow \rho_2 \, second\_moment + (1-\rho_2)\widehat{grad} \odot \widehat{grad} \tag{2.18}$$

, where ρ_1, ρ_2 are the exponential decay rates for the first and second moments respectively. The correction biases of the first-order momentum and the second-order momentum are given by:

$$\widehat{first\_moment} \leftarrow \frac{first\_moment}{1-\rho_1^{ts}} \tag{2.19}$$

$$\widehat{second\_moment} \leftarrow \frac{second\_moment}{1-\rho_2^{ts}} \tag{2.20}$$

, where ts is the time step.

2.3.3 Recurrent Neural Networks (RNNs)

In this subsection, an overview of the basic RNN model and long short-term memory.

Basic RNN

RNNs are able to to leverage the temporal data by handling the sequential aspect of the inputs. So, RNNs exploit the past history of the data by presenting recurrent connections between neurons, that is, that past information from previous time steps. Thereby, they are capable of modelling the temporal dynamics of the input. In a classification task, each input vector is classified by presenting the past temporal context. The length of this context is automatically learned through the weights associated to the recurrent connections. However, the big drawback of RNNs to classify sequences is practically restricted to only few time

steps due to the vanishing gradient problem [72]. For simplicity, the back-propagated error is exponentially dependent on the magnitude of the weights and this leads the error either to vanish or to blow-up. Vanishing gradients means that the gradients is becoming zero or very close to zero which lead to the fact that the weights of the networks is not going to be changed [54]. Blowing-up problem is the opposite which the gradients is becoming bigger and bigger and then it will not converge. Consequently, the weights become constant or oscillating and the network will then be unsuccessful to learn the long-term dependencies [54]. That means that the long time lags can not be accessible in the existing architectures. The input of the network is defined by 2.21 and hidden layer output is defined by the equation 2.22 and the final output is given by 2.23.

$$y_\tau^{(0)} = x_\tau \tag{2.21}$$

$$y_\tau^{(l)} = f^{(l)}(W^{(l-1,l)} y_\tau^{(l-1)} + W^{(l,l)} y_{\tau-1}^{(l)} + b^{(l)}) \tag{2.22}$$

$$y_\tau = f^{(L+1)}(W^{(L,L+1)} y_\tau^{(L)} + b^{(L+1)}) \tag{2.23}$$

for each layer l = 1, 2, ..., L and time step τ = 1, 2, ..., T where T is the number of time steps. The input layer corresponds l = 0 and the out layer corresponds to $l = L+1$. Here, $y_\tau^{(l)}$ refers to the hidden layer. Likewise, $W^{(l-1,l)}$ is the weight matrix representing the feed-forward weights from layer $l-1$ to layer l whereas $W^{(l,l)}$ represents the recurrent weights of the layer l. The bias vector is denoted as $b^{(l)}$ and $b^{(l)}$ is the activation function for layer l. x_τ is the input to the network and y_τ is the final output.

2.3.4 Long Short Term Memory

The Long Short Term Memory Network (LSTM) [73] was inspired by the analysis of the vanishing gradient issue in RNN. An LSTM layer is composed of recurrent connected blocks, defined as memory blocks rather than simple neurons as in the hidden layer of the MLP. Each block contains one or more recurrent connected cells in addition to three multiplicative units. The multiplicative units are the input, forget and output gates. The gates are the activation of non-linear summation from inside and outside the block and controls 'remember' or 'overwrite' operations of the cell. The gates commonly use the sigmoid function. Hence, the output of the gates is a probability $[0,1]$. Intuitively, the memory cell remembers the previous input as long as the forget gate is open and the input gate is closed. More precisely, the input, forget and output of the memory cell is multiplied by the activation of the input, forget and

output gates respectively [59]. The LSTM cell is displayed in Fig. 2.4. While the network task is classification, the cells contents is ruled via the gates.

$$i_\tau^{(l)} = \sigma(W_{h,i}^{(l-1,l)} y_\tau^{(l-1)} + W_{h,i}^{(l,l)} y_{\tau-1}^{(l)} + W_{c,i}^{(l,l)} C_{\tau-1}^{(l)} + b_i^{(l)}) \tag{2.24}$$

$$f_\tau^{(l)} = \sigma(W_{h,f}^{(l-1,l)} y_\tau^{(l-1)} + W_{h,f}^{(l,l)} y_{\tau-1}^{(l)} + W_{c,f}^{(l,l)} C_{\tau-1}^{(l)} + b_f^{(l)}) \tag{2.25}$$

$$o_\tau^{(l)} = \sigma(W_{h,o}^{(l-1,l)} y_\tau^{(l-1)} + W_{h,o}^{(l,l)} y_{\tau-1}^{(l)} + W_{c,o}^{(l,l)} C_{\tau-1}^{(l)} + b_o^{(l)}) \tag{2.26}$$

$$C_\tau^{(l)} = fg_\tau^{(l)} \odot C_{\tau-1}^{(l)} + ig_\tau^{(l)} \odot tanh(W_{h,c}^{(l-1,l)} y_\tau^{(l-1)} + W_{h,c}^{(l,l)} y_{\tau-1}^{(l)} + b_c^{(l)}) \tag{2.27}$$

$$y_\tau^{(l)} = og_\tau^{(l)} \odot tanh(C_\tau^{(l)}) \tag{2.28}$$

where $\odot$ represents the element-wise product, $\sigma(.)$ and tanh(.) are the element-wise sigmoid and hyperbolic tangent function respectively. $i_\tau^{(l)}$, $f_\tau^{(l)}$, $o_\tau^{(l)}$ are the input, forget and output gates respectively. $C_\tau^{(l)}$ is the memory cell activation. Note that the weight matrices $W_{c,.}^{(l,l)}$ are diagonal, different and are from cell to multiplicative gate.

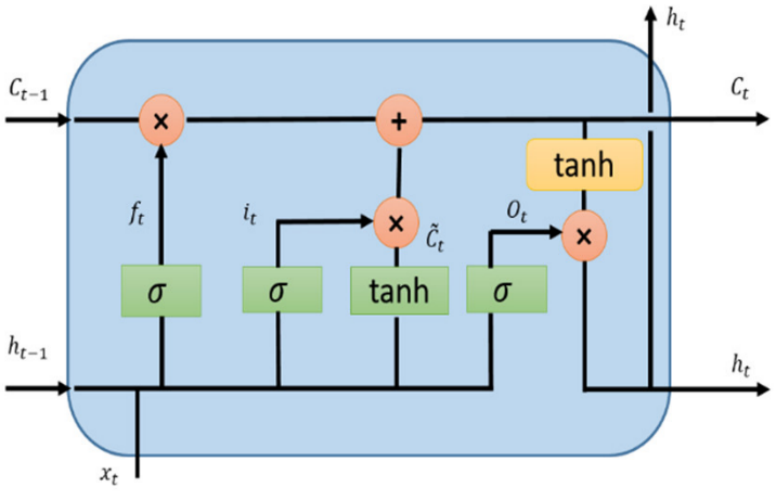

Fig. 2.4 LSTM Cell [4]

In order to understand how the gates in LSTM works, Fig. 2.5 is presented for that purpose. I used *O* to refer to the *open* state and − to refer to the close state. Suppose we have a time series with 6 timesteps (small number to be clear in the visualisation). The time-series data is fed to the network (one by one timestep) from the input layer to the hidden

and according to the the probabilities of the input gate, the network decide to accept the input and overwrite the state of the cell in the hidden layer. The timestep 1 overwrites the state in the hidden layer where the input gate is open and since the output is closed it did not change the state in the output layer. The forgot gate for timestep 1 is also open, which means it will be used again in the next timestep (the network will remember it). Then for timestep 2, the input gate is closed and the forget gate is open and therefore the previous state will be used and the new input will be ignored. Likewise, the output gate is open in timestep 2, that means the state from hidden layer will overwrite the output layer. From the figure, we notice that the network is remembering the timesteps from 1 to 5 and not accepting the inputs. Nevertheless, in the timestep 6, the input gate is open which means the network will overwrite the hidden block and forget the past history. Also, the output gate is open and then the state will be transferred to output block.

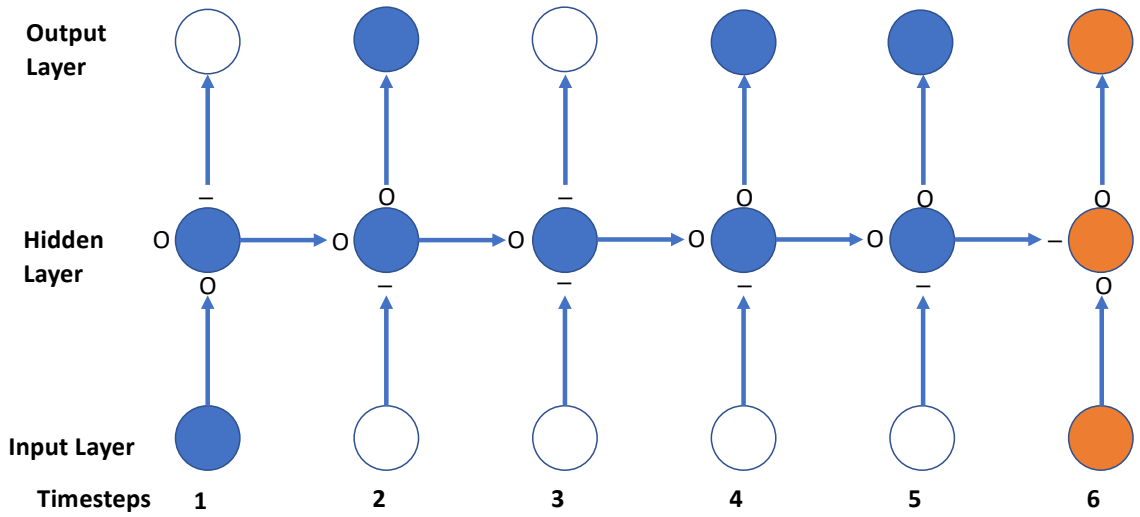

Fig. 2.5 LSTM Example. The gates are illustrated in hidden states where the bottom gate represents the input gate, left one represents the forget gate and the top represents the output gate. The figure shows how the timestep 1 carried forward to the next hidden states or the output state. Since the 1^{st} input gate is open, the state is carried forward to the 1^{st} hidden and not to the output state since the output gate is closed. The forget gate is open, that means the state is carried forward to the next hidden state. Note that the hidden state do not take the 2^{nd} timestep since the 2^{nd} input gate is closed.

2.3.5 Convolutional Neural Networks

Convolutional neural networks (CNNs) were originally presented by K. Fukushima in 1988 [48]. CNNs [96] are a particular kind of ANN for analysing topologically grid-like data.

Due to the advances in Computation hardware, CNNs became widely used in the last decade. CNNs outperforms DNNs in several ways, one of these ways is that CNNs act like as a visual processing system of a human by being capable of extracting the abstraction of the two dimensional features and more effective at learning process. Hence, CNNs does not need a pre-training step for feature extraction but it can achieve it in an embedded manner. The pre-training step can be achieved by deep belief networks or autoencoders to reconstruct the input by these simple layers in order to be utilised for feature extraction and then the output of these layer is considered to be the input to the network [54]. However, CNNs don't need this at all and have the ability of achieving feature engineering within the convolutional layers. CNNs comprises two main processes: the feature extraction learning process followed by the classification process. The layers which are responsible for the feature extraction learning process are convolution layer and subsampling (pooling) layer. A Convolution is a mathematical operation on two functions that produces a result function expressing how the shape of one function is modified by the other. Hence, CNN is using convolution as an alternative to matrix multiplication. The convolution process can be defined as:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(a)g(t-a)da \tag{2.29}$$

The convolution in Eq. 2.29 can be described as a weighted average of the function f and g is considered as the weighting function. The function f is interpreted as the input data, g is the kernel and the output (after applying the activation function) of the convolution is interpreted as the feature map. The motivation beyond CNN is that its capability of sparse connectivity (sparse weights), parameter sharing, equivariant representations and handling variable size inputs. Sparse connectivity can be achieved by choosing a kernel to have smaller size than the input itself. Parameter sharing refers to using the same parameter (weights) for more than one function in a model whereas in MLP the weights are used once. The idea is that some weights are revisited when the kernel is passing over the input. Therefore, the same weight can be shared for multiple kernels. The way that the CNN shares its parameters create an equivariant translation. That means when the input changes in some way, the output changes in the same way [54]. In Fig. 2.6, an illustration of of the CNN parameter sharing and how the neurons are connected which differs from the MLP structure. The figure illustrates the weight sharing within the neurons, for instance the centred wight (the back arrow) from the node (x_2) is shared over the nodes (s_1, s_2, s_3)

The pooling layer provides a basic statistics on the output of the activations from the convolution layer such as (maximum or average) and max-pooling is the commonly used and widely applicable by researchers recently. The output nodes from the convolution layer and pooling layer is called feature maps. The feature maps are propagated from lower layers

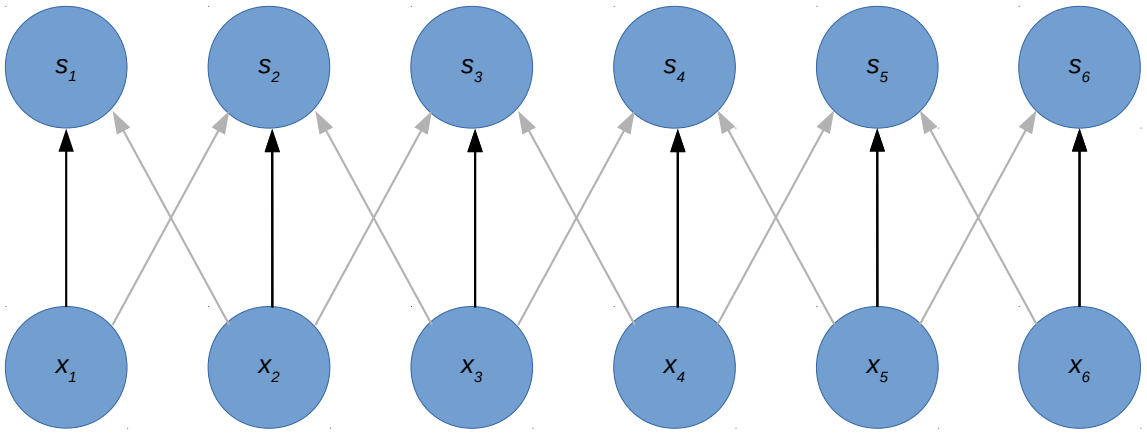

Fig. 2.6 This example shows the parameter sharing. The black arrows represent the middle neuron of a 3-neuron kernel. Due to the parameter sharing, all input locations will use the same parameter.

to higher layers constituting high-level features [54]. Within the propagation of the feature maps through the layers of the network, the dimensions of the features are shrinking based on the kernel size and the stride (step size) for the convolution process and also on the kernel size for the pooling process. Nevertheless, the number of feature maps is often risen up to assure a better presentation that could help later in the classification layers. The output activations of the last layer of convolution is the input to the fully-connected layers to perform as a classifier. Feed-forward neural networks (or simply dense layers or fully-connected layers) is widely utilised as the classifier layer due to their classification performance [116]. The basic structure of the CNNs networks is presented in Fig. 2.7.

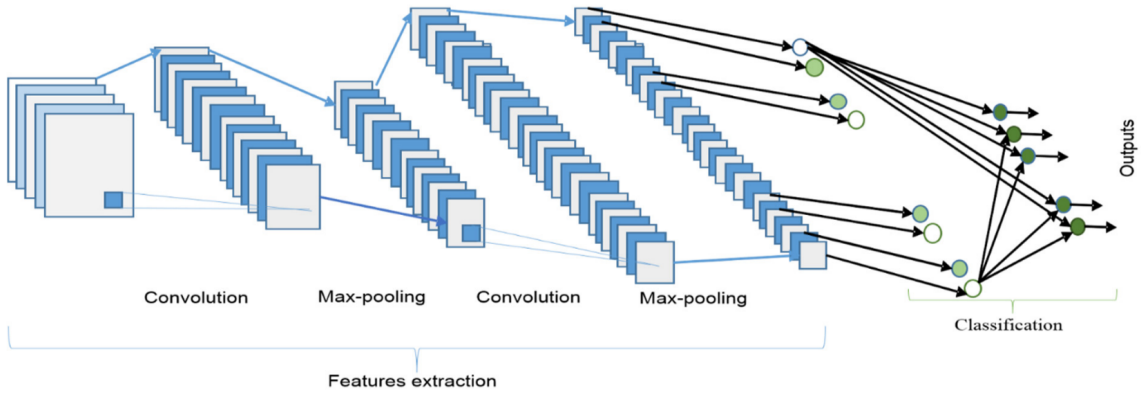

Fig. 2.7 Basic Structure of CNNs [4]

There are different versions of CNNs: 1D, 2D, 3D. The convolutional 1D is appropriate for the time-series data and this is the main focus of this thesis whereas 2D is suitable for images and 3D is often utilised for instance in video data or 3D images like MRI in the medical imaging. In the convolutional 1D, the feature maps are convolved through the time-steps of the time-series sequence with the learning kernel that moves over the time-steps depending on step size that set the overlap for the convolution process. The output of these kernels is fed to a linear or non-linear activation such as sigmoid, rectified linear unit (ReLU), identity function or hyperbolic tangent. The most used activation function in the convolution process is the ReLU. Then the output of this activation function is passed through to the subsampling (pooling) layer. The subsampling layer downsamples the input time-steps according to the size of the kernel of the subsampling layer. However, there is no change in the number of both input and output feature maps [4]. For instance, if the time series (sequence) length is n time-steps and the kernel size for the pooling layer is 2 and the average pooling function is applied, then the output time-steps is the average of each two consecutive time-steps. So, the output will be halved.

The fully-connected layer then receives the final feature maps represented as vectors of scalar values and then applies the softmax activation. In the backpropagation operation in CNNs, the fully connected layer is updating its parameters using the normal backpropagation updates for feed-forward networks. Notwithstanding, the updates of the filters in the convolutional layers are obtained by the convolution process between the current convolutional layer and the previous layer not by backpropagation like in the fully connected layer.

One of the most important concepts for CNN is the *receptive field*. The receptive field can be considered as the region in the input that influences a particular CNN's unit (neuron). This unit not only focuses on the receptive field as a region but focus more on the centre of it. So, that means in case of CNN 1D the time-step close to the centre of the receptive field contributes more to the computation of the output unit. The receptive field size relies on the kernel size [4]. The receptive field is very important to handle the temporal variation. The number of parameters in CNN model or any deep learning model generally sets the complexity of the computational model. In order to calculate the size of the output units, the following equation [42] is formulated for that purpose:

$$n_{out} = \left(\frac{n_{in} - k + 2p}{s} \right) + 1 \tag{2.30}$$

where, n_{out} is the number of dimensions of the output, n_{in} is the number of dimensions of the input, s is the step size (stride), k is the kernel size and p is the padding size. In order to obtain the receptive field size, this depends on the kernel size and on the dilation size

(distance between two consecutive features) as in the following two equations [42]:

$$d_{out} = d_{in} * s \tag{2.31}$$

$$r_{out} = r_{in} + (k-1) * d_{in} \tag{2.32}$$

where, d is the dilation size and r is the receptive field size and starts with 1 in the beginning. Note that if the $s \neq 1$, then d_{in} must equal 1. In order also to obtain the centre coordinate of the first feature, the following equation [42] is formulated:

$$c_{out} = c_{in} + \left(\frac{k-1}{2} - p \right) * d_{in} \tag{2.33}$$

Note that c_{in} starts with 0.5. An easy way to increase the receptive field of output units without increasing the kernel size is by exploiting dilated convolutions which is especially effective when multiple dilated convolutions are stacked one after another like in WaveNet model proposed by Oord et al 2016 [123].This model basically leverages an autoregressive generative model that utilises the dilated convolutions on the raw audio to condition new frames on a large context of past frames.

2.3.6 Autoencoders

An autoencoder [95],[69] is ANN that is trained to attempt to reconstruct its input. Autoencoder comprises an encoder function that codes the input and produce a latent representation and a decoder function that produces a reconstruction from the latent representation. The general idea of autoencoders is the stochastic mappings for both of the encoder and decoder functions: $p_{encoder}(h|x)$, $p_{decoder}(x|h)$. Autoencoders are trained using recirculation and backpropagation whereas MLP is trained using backpropagation [67]. Autoencoder model is illustrated in Fig. 2.8. As shown in the figure that the input through the encoder layer which compress the input until the bottleneck layer and then the decoder tries to decode the last encoded output back again to reconstruct the input. Examples of autoencoders include sparse, autoencoders, denoising autoencoders or convolutional autoencoders, among others. Convolutional autoencoders are discussed in details in Chapter 4.

A sparse autoencoder is a variant of autoencoder which add a sparse penalty in the loss function in addition to its reconstruction error. Specifically, the penalty ($\propto$) is added to the encoder layer and this is illustrated in the equation 2.34.

$$L(x; g(f(x))) + \propto (encoder) \tag{2.34}$$

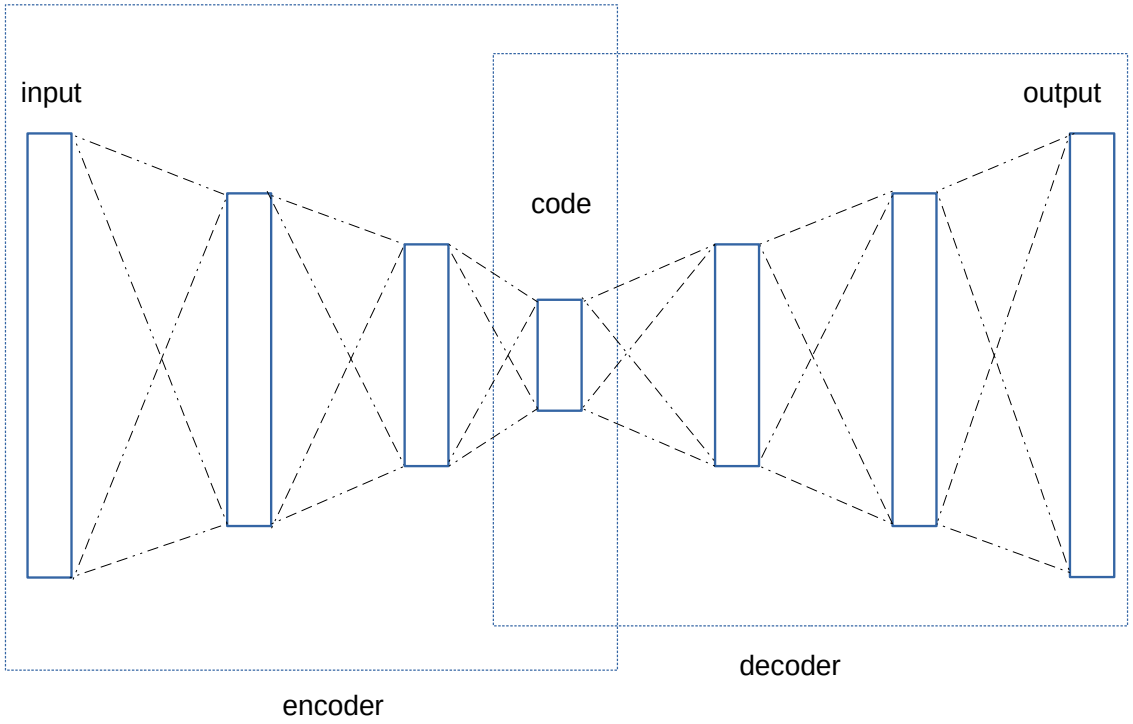

Fig. 2.8 Autoencoder Structure

where, $encoder = f(x)$ is encoder output, $g(encoder)$ is the decoder output [54].

Classification tasks usually uses the sparse autoencoder to learn and extract features from the raw data. The sparsity condition as a regulariser to the autoencoder is sensitive to some unique statistical characteristics within the training that helps the model to learn useful features not just act as an identity function. Therefore, the sparse autoencoder can be described as a maximum likelihood approximation of a generative model [134].

Unlike sparse autoencoders, denoising autoencoders receive a corrupted input with noise added to it and the encoder attempt to address identity-function risk in order to denoise the corrupted input (i.e. reconstruct the original input from the corrupted one). The corruption level is expressed as a percentage of the input $(0,1)$. The reconstruction can be accomplished by maximising the likelihood or simply minimising the negative log-likelihood. As an alternative to the maximum likelihood, Hyvarinen [77] proposed score matching approach. The idea of score matching approach is that at each data sample, the model is attempting to have the same score as the data distribution.

2.4 Human Activity Recognition

The early study on Human Activity Recognition was first done by [1]. HAR research started from considering only videos and images and then extended to the utilisation of wearables and ambient sensors [190]. In this section, the main focus is only the wearable sensor-based HAR systems as this is the main focus of the thesis. Specifically, a background on HAR systems and some details for each step of these systems will be introduced. This background includes an overview on sensors used in HAR, introduction to speech activity detection, introduction to daily life activities, pre-processing of the data, feature extraction, classification methods.

2.4.1 Introduction to HAR

The evolving of wearable devices that includes smart-phones, smart watches, wristbands allows for continuous recording and unobtrusive monitoring of the human speech or physical activities. Speech activity detection (SAD), also referred to as Voice activity detection (VAD) has recently received a great deal of attention and is a key component for many speech processing applications [99]. SAD will be used in this thesis. SAD is used to identify speech segment boundaries from observed audio data. Most speech processing applications rely on speech detection as the first preprocessing stage such as speech coding, automatic speech recognition (ASR) and speaker verification [99].

The second aspect of HAR is the physical activities such as walking, running, cycling and sleep patterns [45],[62]. The idea of activity recognition is identifying the patterns of specific activities based on some knowledge for each activity such as signal vector magnitude and metabolic equivalent task (MET) [62]. The application of activity recognition in healthcare varies from Depression which is the main focus of this thesis, rehabilitation after stroke [94] and analysing the mobility conditions of Parkinson's [94].

The flowcharts of HAR systems is presented in Fig. 2.9. The flowchart starts with the raw data which is recorded by various wearable sensors such as tri-axial accelerometer, gyroscope and heart-rate sensor. This raw data is then collected online or offline for pre-processing step. The pre-processing involves the filtering, splitting the data into fixed-length frames and normalization. Following that is feature extraction which includes the basic statistics, time domain and frequency domain features such as mean, standard deviation, dominant frequency, entropy, signal vector magnitude, pitch and fast Fourier transform coefficients. These features are then reduced by dimensionality reduction approach or feature selection algorithm in order to get a smaller compressed set of the best features. This reduction helps for further learning and alleviate the issue of computation. The final set of best features is then fed to the training algorithm for classification. This is typically the full process of

classification using classical machine learning algorithm. The deep learning (DL) based classifiers (which is the second part of the flowchart) does not need the whole pre-processing procedure. It only needs the data to be normalised. Then the normalised data will be fed directly to the deep learning model and the deep learning model will take care of extracting the features automatically throughout the layers of the deep network. The deep networks includes but not limited to RBM, Deep Belief Networks DBN, Multi-layer perceptron, RNN, CNN, Autoencoders. These networks are able to perform the feature learning and learning the model at the same time. All these methods are discussed in the following subsections.

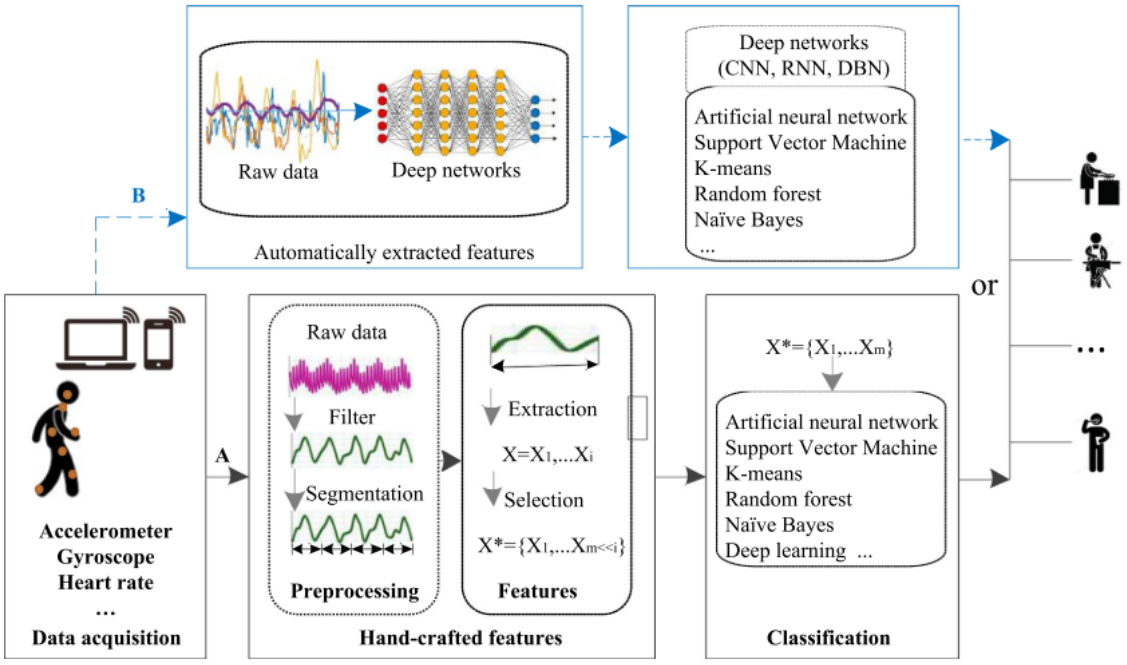

Fig. 2.9 HAR systems flowchart [175]

2.4.2 Wearable Sensors in HAR

Wearable sensors have recently gained the potential to provide assisted living in healthcare applications and well-being. The design of these wearables requires small size, lower battery consumption, small size and high accuracy. The wearable sensors types include inertial sensors, physical health sensors and environmental sensors. The inertial sensors such as microphone, tri-axial accelerometer, gyroscope and magnetometer are able to measure the acceleration, angular velocity, magnetic fields, vibration, rotation and shock [62],[25]. In [60], Guo et al. build a mobile app for patients' activity recognition using the built-in tri-axial accelerometer, gyroscope and magnetometer. The demanding issue with inertial sensors is

the battery life and the calibration required in order to get accurate measurements. Physical activity sensors such as Electrocardiogram (ECG), skin temperature and heart rate have been utilised often with the inertial sensors to capture critical signals especially in the healthcare field. Temperature, humidity, light sensor, barometer, etc. are examples for environmental sensors. In addition to all the previously stated sensors, some researchers are using the camera, microphone, GPS, etc. [175]. However, these other sensors are not that common in the physical activity analysis.

The platforms that the wearable sensors can be presented in such as smart watches, smartphones, smart clothes, inertial measurement unit and specifically designed hardware required for a specific task. Smartphones nowadays are advancing rapidly and acquire a wide range of sensors, memory, GPS and battery that increase the capability of HAR system where the data from the sensors can be collected easily [62]. Smart watches are basically wrist mounted devices and are substantially designed to integrate with a phone or a computer. That makes them more convenient than smartphones and also unobtrusive due to the relative body location and they do not need to be carried all the time like smartphones. Smart clothes are able to achieve better measurements since they can include more than one sensor. Smart clothes have the advantage of easy wearing and therefore they are capable of long-term monitoring [2]. Inertial measurement unit is a special hardware that measures and detect a combination of factors previously mentioned. Inertial measurement unit can be used to detect human gestures or physical activities [113]. Specifically designed platforms are to achieve special objective with specific hardware such as in [167] for recognising the hand washing and drinking where the authors used nine-axis device to achieve this task. Another example for specifically designed platform is a wearable eight-channel bio-potential data collection platform. This platform is connected with tri-axial accelerometer and ECG and potentially used for healthcare related information [30].

Sensor placement refers to the body locations that the sensors can be placed. The sensor location can be various along with different applications. For example, a foot-mounted accelerometer are able to measure acceleration of the foot and hence it has the potential to be used for gait analysis or step count [5]. Normal activities, such as brushing teeth, cooking and ironing are meant to be acquired based on wrist-worn-sensors [24]. Other activities (leg-involved activities) which comprise walking, running, jogging, riding, etc. incorporate the thigh-located sensors [112]. The placements can be categorised into four categories: one-to-one, one-to-many, many-to-one and many-to-many. One-to-one [159] means one sensor is placed on a single body location. However, this might lack sufficient information required to obtain better measurement where the single body location is not enough to provide rich information and also in case only one sensor is used. The second category is one-to-many

which refers to single sensor on multiple body locations in order to get rich information from different body parts. Many-to-one aims to measure a variety of information from multiple sensors and one body part. This category is better than one-to-many in terms of the diverse source information obtained from various sensors. The final category is many-to-many; in this category there are multiple devices that are embedded with multiple sensors on multiple body locations[25].

2.4.3 Data Preprocessing

The raw data required to be preprocessed before feeding to the classifier. The common preprocessing approaches are such as window framing (segmentation), filtering and normalisation. Filtering is a process of cleaning the raw signal to remove unwanted artefacts and noise. The filtering is commonly used before the window segmentation prepared for feature learning and extraction. There are various filtering methods, low-pass filter, second-order Butterworth High-Pass filter, median filter and N-point moving-average filters[79],[104]. On the contrary to filtering, some researchers believe that keeping the signal without filtering is better especially that will not lead to any loss of any information that might be rich for classification.

The time series data coming out from the wearable sensors are relatively large and need to be split up to smaller segment to learn the important components from them and that is called window segmentation. That relies on the sampling rate of these sensors. The sampling rate for HAR is usually between 20Hz to 100Hz. The common window segmentation approach is the sliding window which splits the time series into fixed-size frames with fixed-size overlap. The commonly used overlap between two consecutive frames is usually 50% and this overlap is to reduce the information loss at the edges of each frame [35].

The frame length is varied from 1s [15], 2.56 s [62] , 5s [104] to 30s [102] and sometimes exceeds half a second and reach multiple seconds. The small frame size makes it easier and rapid for later use in extracting the features. However, these small frames might not contain less than even one cycle of any activity. On the other hand, the large frame size can contain multiple cycles of one activity and probably more than one activity which means the frame becomes full of useful information but this is does not guarantee enhancing the performance but even make the recognition learning delayed. Some researchers do a grid search by choosing a range of frame lengths and then choose the best but the best is not applicable to any other tasks where is the optimal one is depending on the case and the application. In order to make easy for choosing the frame length, [75] set some conditions for that. The first condition is that the frame contains at least one cycle of the activity. The

second condition recommends the frame length to be 2^n so it can be easily utilised in the frequency domain i.e. when applying Fourier transform [173].

2.4.4 Feature Learning

A feature is an individual property that can be measured or a characteristic of a phenomenon that can be observed [16] and it is the input for most machine learning classifiers. Feature extraction from the raw time series data can be achieved by either manually calculated (hand-crafted) or automated learning. Hand-crafted features require domain knowledge and calculated for each frame either in time domain or frequency domain. The features objective is to highlight the latent representation of the data in order to be used discriminate between activities in HAR. Hand-crafted features is categorised into: time-domain, frequency domain and hybrid features.

Speech Feature Extraction

Many algorithms have been proposed to create robust SAD models. These algorithms are exploiting the feature extraction engineering such as extracting the time-domain features (Signal energy based SAD) [88, 92, 23]. The time domain features is a good option for very small datasets and in a clean environment. However, these features are not enough to learn from the data. Looking at the frequency domain features can be an alternative to this issue such as LPC residual domain [117], mel-frequency cepstral Coefficients (MFCC) [51], short-term spectral features [65] and long-term frequency components [50] which they depend on Fourier transform to transform the data from time domain and frequency domain. So, it performs better than the time domain as in [188]. The only issue with these features is that its sensitivity to the frame length and the frequency band chosen. Signal high-order statistics is also an approach to extract the features from the data such as in [103] and modulation spectrum analysis [80]. Apart from the time domain and frequency domain features, a fusion of acoustic features (time and frequency domains) with multiple observations or joint spectra-temporal features is a better option according to [166, 188]. The only issue of the combination of features from different domains is that some features become redundant and then it require feature selection or feature reduction approaches which is time consuming and need extensive computation. All of these models lack some latent representations that can be learned from the original signal itself [58]. Therefore, deep learning models have emerged to perform an efficient feature engineering as part of its processes. Building robust SAD models is a big challenge especially if the speech is recorded in highly unconstrained naturalistic, ambient environments that comprise various additive noises and distortions.

Physical Activity Features

Time domain features are statistical measures that can be calculated directly from the frame samples. For example, if we have acceleration and we want to discriminate between static and dynamic ADL, then the signal vector magnitude for each frame should be able to produce different signal strengths. Therefore, we can classify the static and dynamic activities (such as sitting and walking). This is just a very simple case for the data but when it comes to real world data, then more efficient features will be necessary. The authors in [104] and [62] use signal magnitude area (SMA) fused with other features to improve the recognition accuracy of dynamic activities. Standard deviation (Std) achieve consistently high accuracy to classify activities such as standing, walking and climbing stairs according to [94]. Some other researchers believe that median, variance, skewness, zero crossing rate, autoregressive coefficients , peak-to-peak and so on can achieve better classification accuracy in HAR [104], [160],[62].

Frequency-domain features represents the periodicity of signals i.e. how much of the signal lies in frequency band over a range of frequencies. In order to obtain frequency domain features, a transformation should be applied to the frames such as Fast Fourier Transform (FFT), Discrete Wavelet Transform (DWT), or Discrete Cosine Transform (DCT) [175]. Frequency domain features with FFT as a transformation comprise spectral energy, entropy [62], dominant frequency[173],[160]. Alickovic et al. [3] adopt wavelet packet decomposition (WPD) for extracting the frequency domain features from EEG signals. Other studies employed the DCT for their feature extraction and achieved high accuracy (97.51%) for their proposed automated cardiac arrhythmia detection. On the other hand, the hybrid features are extracted from multiple channels of a sensor or multiple sensors i.e. fusion of sensory features. E.g. Examples of fusion of features includes tilt, rotation and yaw which represents the attitude of the wearable sensors. In order to obtain these hybrid features, multi-channel sensor required or a set of inertial sensors such a tri-axial accelerometer, a gyroscope or a magnetometer. In [160] the user posture orientation was detected by exploiting the tilt angle.

On the contrary to manually or hand-crafted features, deep features (automated learned features) are considered to better feature extraction. Automated learned features are features representation methods that uses deep learning (which will be discussed in a separate section). Deep learning provides the capability of learning the features from raw sensor data in automated technique that optimise the hyper parameters layer by layer starting from encoding the input followed by the decoded output should reconstruct the input [173]. Deep features are exploited to improve the recognition performance of ADL [61] where the authors use deep belief networks to extract the features. Another study deep convolutional neural

network (CNN) [136] and they provide a thorough analysis of the effect of extracting the features from different layers on increasing the number of feature maps. Panwar et al. [126] proposed a CNN model to classify the movement of the fore-arm of people performing ADL using a wrist-worn accelerometer.

2.4.5 Dimensionality Reduction and Feature Selection

Having more features especially hand-crafted, frequency domain or hybrid features will provide more information which in turn should help the classifier and lead it to better performance. However, the initial group of features can be redundant or require a higher computation cost. That will lead to low learning efficiency and therefore the model is overfitting. Overfitting means that the model learns all the details of the training set and therefore will deal with the noise in the data as concept. So, this will affect the model performance negatively on the test data which is not seen before by the model making it not possible to generalise [14]. In order to circumvent this issue, features dimensionality reduction and feature selection are considered.

Feature dimensionality reduction algorithms is generating linear or non-linear combinations of the input in an unsupervised manner to reconstructs the original features. The common dimensionality reduction algorithms are principal component analysis (PCA) [63] and Kennel PCA [62]. PCA is a linear transformation of the original features to a new feature space and then find the best projection depending on the variance. The output of PCA is the principal components. The best principal components are the components with high variance. Kernel PCA on the other hand is a variant of PCA but use a non-linear transformation of data by mapping the features to higher dimensional space using kernel function (e.g., radial basis function) then the follow the same procedure of PCA [180]. The issue with PCA is that it did not take the class labels into consideration since PCA is an unsupervised learning method. This problem can be tackled by Linear Discriminant Analysis (LDA) but in this case it converts the original feature space to lower dimension to take into account the class separability [168]. The non-linear variant of LDA is Kernel LDA which is used in HAR like in [62].

In addition to the classical dimensionality reduction algorithms, deep learning can also be used for the same purpose. One of the most popular approaches is using the autoencoder which aims at learning a representation from the input features in order to reconstruct the input by minimising a loss function (such as mean square error, mean absolute error or cross entropy). Recently, the autoencoders become a promising approach for different applications including the applications of time series analysis like HAR by providing a meaningful

information from the features input or even from the raw data itself [19]. The details about autoencoders and how it works will be presented in the Deep Learning section.

Feature selection usually utilise an optimisation method to find the best set of features that help more in the training and perform well in the recognition process without mapping or projecting the original features to a different space. There are various methods for feature selection. These methods usually are to rank the features according to some conditions such as similarity measures, distance function, correlation and consistency and this process is separate from the classifier. One common example is mutual information based feature selection. This method fundamentally relies on the correlation between features [8]. Another set of feature selection methods exploits some classifiers to evaluate the feature efficacy on the final performance and then rank the feature according to their relevance. The process starts with a subset of the features followed by the performance evaluation and then adding more features until it gets to the best performance or accuracy. The process is utilising grid search. One conventional example for these methods use support vector machine [10]. The issue with this method is that it has a very high computation cost. In addition to these methods there is another type of feature selection methods called embedded methods. This integrates any of the previously stated feature selection method by adding regularisers into the optimisation algorithm for the model training. The most conventional embedded approaches such as Lasso [98] and ridge regression [101]. Lasso is a method of applying l_1 L1 norm regularisation to the training optimisation and choose the features with high weights whereas ridge regression is applying l_2 regularisation.

2.4.6 Classical Machine Learning based classification for HAR

Classical Machine Leaning Classification for Speech

Several speech processing systems rely substantially on SAD. The efficacy of speech communication systems can be improved by detecting only the speech segments/frames. Furthermore, SAD models enhance the performance of speech recognition systems [32] by filtering out the segments of background noise in addition to the unwanted silence. One of the obstacles/limitations for SAD is to make it perform well in low signal-to-noise ratio (SNR) environments such as naturalistic/unconstrained environment settings (e.g. street recordings with background car noise). Machine learning based SADs have received a great deal of attention due to the integration of systems of speech recognition and the advantages of multiple audio features fusions. The machine learning-based SAD systems can be classified into supervised and unsupervised learning. Many unsupervised approaches rely basically on dimension reduction methods such as principal component analysis [139], non-negative

matrix factorization [163] and k-means clustering [57]. Some others have been developed for detecting speech using statistical methods [139], [141].

On the other hand, Supervised learning based SADs can be divided into six categories. The first category is the discriminative weighted training based SADs [157]. They exploit the linear weighted combinations of a set of features in the original feature space. The second category is the kernel based methods such as support vector machines (SVM) [182]. In this category, multiple features are combined/fused to form a long feature vector in the original feature space, and then this long feature vector is projected on the kernel induced feature space for better performance of classification predictions. The third category is the multiple-kernel SVMs based SAD [181]. This category makes use of the diversity of multiple features by projecting them into different kernel spaces and then combines these features in the kernel spaces in such a way with linear weighted combination. All the above mentioned three categories utilize the shallow models, i.e. these models are with only zero or one hidden layer, which lack the ability of describing highly variant features and discovering the underlying manifold of the features. The fourth category is probabilistic models such as Gaussian mixture models (GMM) [47] and hidden Markov models (HMM) [170]. The fifth category is ensemble methods such as random forest [150]. In this category, multiple learning algorithms are combined in order to obtain better predictive performance than could be obtained from any of these learning models separately. The last category is the deep learning models which are the main focus of this thesis.

Classical Machine Learning for Activity Recognition

After the features have been extracted from the raw data and then processed, they are fed to the classifiers to build the classes from the data. The approaches for classification can be divided to two types: classical machine learning based and deep learning based classification. There are various algorithms for supervised learning that have been applied to HAR such as SVM [109], Decision trees [114], k-Nearest Neighbours (kNN) [2], Random forest (RF) [127], etc. Other research use Baysian classifier with kNN to investigate the outliers effect on the whole model [6] where the authors presents HAR framework assessing different placement of the sensors and their corresponding efficacy. Mehrang et al. [109] present a single wrist-worn device that include tri-axial accelerometer and a heart-rate sensor for HAR to classify basic activities like standing, sitting, stationary, cycling, household and they exploit RF and SVM apply RF and SVM to achieve this task. All of these classifiers are very sensitive to the input features and the dimensionality reduction applied to it after that which is an issue that puts these methods at risk to give a good activity recognition accuracy. Some studies fuse different classifiers to circumvent this issue like in [26] where the authors use

a fusion of MLP, SVM and genetic algorithm for optimisation and to investigate the best combination.

In unsupervised learning for physical activities, the features of each activity are passed to the model and then each activity will have a separate model corresponds to it. The traditional unsupervised learning method is cluster analysis that aims to extract the latent representation from the features to group the input data into clusters. These clusters use a similarity function like Euclidean distance function or any other similarity measure to find the decide the correlation between the points in the same cluster. Examples of unsupervised learning include but not limited to k-Means, Gaussian mixture models (GMM) [91] and Hidden Markov models (HMM) [169].

2.4.7 Application to HAR: RF and HMM for Activity Recognition

As discussed in Section 2.2, RF cannot handle the temporal variation of the data. So, in order to achieve this purpose, hidden Markov models (HMM) [132] are leveraged in this work where HMM encodes the temporal variation of the sequence of the classes and consequently help make better prediction. HMM comprises hidden discrete state state space. The sequence of these states has the Markov property. The Markov property means that the probability of a state at a timestep t given the present and past history of state at the previous time-steps is the probability of the state at timestep t given only the state at the previous timestep i.e. the future is only influenced by the present and does not depend on the past. If there is a sequence of states $s = s_1, s_2, ..., s_{t-1}, s_t, s_{t+1}, ...$, then the probability is given by Eq. 2.35:

$$p(s_t|s_1, s_2, ..., s_{t-1}) = p(s_t|s_{t-1}) \tag{2.35}$$

The reason that HMM is called hidden, that is due to the fact that HMM is looking at both of the observed events and hidden states (s_t) simultaneously. For each timestep t, the hidden states s_t is assigned one of the classes. So, if there n classes ($c_1, c_2, ..., c_n$), then the transition matrix from class i to class j is computed from the training data and can be given by Eq. 2.36:

$$T_{ij} = p(s_t = c_j|s_{t-1} = c_i) \tag{2.36}$$

The hidden states s_t have stochastic emissions e_t that rely on them. The stochastic emissions are sampled from the probability distribution $p(e_t|s_t, \Theta)$, where Θ are the probability distribution parameters. These emissions are computed by the votes from the trees in random forest. In the preprocessed movement data, one can think of the hidden states sequence s_i as the ground-truth physical activities whilst the stochastic emis-

sions can be seen as the predicted ones. The full pre-trained model is available at https://github.com/activityMonitoring/biobankAccelerometerAnalysis

2.4.8 Deep Learning based classification

Deep Learning based SAD

Deep learning models have the capability to learn multiple hidden representations from raw data itself [58]. According to the literature, many deep learning based SAD still utilise feature extraction engineering rather than using the raw data. In [76] PLP features have been utilized with RNNs. Further, Log-mel spectrograms features with its deltas and acceleration coefficients are fed to a CNNs as in [166]. Moreover, in [122], Obuchi applied an augmented statistical noise suppression (ASNS) where log-mel filter bank energies features were fed into a decision tree (DT) and a SVM and then to a CNN classifier in order to enhance the detection accuracy. On the contrary, unsupervised learning deep belief networks (DBNs) is used for SAD [188]. Essentially the DBN can generate highly variant features in its multiple hidden layers since the unsupervised pre-training phase of DBN provides initial points that are close to the local optima. However, the deep layers of the DBN-based SAD do not lead to an apparent prediction performance improvement to the shallower layers. Another variant of DBN proposed in [86] for improving the speech detection based on likelihood ratio. In this algorithm, the underlying regularity and manifold of multiple features can be investigated when compared with the shallow model such as SVM and its initial parameters were trained using restricted Boltzmann machines.

Addressing the challenges of DBNs based SAD, de-noising deep neural networks (DDNNs) based SAD [189] has already solved this issue. The DDNN training stage includes two phases: unsupervised de-noising layer-wise pre-training and the supervised tuning. The pre-training process performs a minimisation to the reconstruction loss between the noisy its corresponding clean speech signals in order to obtain the main set of features. The supervised phase is fine-tuning the pre-trained parameters for the minimum classification error. Another variant of DNN is maxout networks with dropout regularisation which has been also proposed for speech detection as in [110]. The idea of maxout units [55] is taking the maximum of the element-wise nodes of multiple fully-connected layers. The superior efficacy of dropout regularisation has been shown on small training data and using maxout-networks, it improves the averaging detection performance of noisy data even under low SNR and in case of training and test data mismatch.

CNNs with their capacity to process structured (1-dimensional i.e. temporal, in this case) data have also been applied to speech recognition. The authors in [166] claimed that CNN

has the capability to learn the spectro-temporal latent representations from the data, although they are not directly aware of the time dimension. RNNs explicitly tackle the temporal condition of speech data frames. RNN automatically leverage the time reliance of the speech samples to their preceding ones. RNNs can distinctly learn the non-linear temporal dynamics of the speech frames in order to make better detection. Hughes and Mierle [76] consider quadratic polynomials as an activation function and claimed that all parameters of the RNN model are optimised together in order to weight its preference for temporal continuity.

Classic RNN architectures have the problems of vanishing gradients and the limitation to handle the data in truncated backpropagation through time (TrBPTT). As discussed in deep learning section, LSTM architectures have emerged to solve these issues. Eyben et al. [44] used basic LSTM with two structures (1 and 3 hidden layers) with 36-dimensional features fed to the network for speech detection with four full-length Hollywood movies as a test set. In [85], Juntae et al. create a learning model for vowel sounds only and not the whole speech segment claiming that the reduced manifold of speech will play an important role in learning more effectively. Furthermore, Leglaive et al. [97] elaborated the bidirectional LSTM for singing voice detection. They applied a method of separating the monaural audio into harmonic or percussive components in order to extract better spectral features and then fed it to the layer of bidirectional LSTM.

Deep Learning for Activity Classification

The issue of classical machine learning is the sensitivity to hand-crafted features. Therefore, DL is utilised for this purpose where DL models have the capability to perform feature learning in an automatic way through the layers of the network. Examples of deep learning classification for HAR involves RNN [61] where the authors incorporate vanilla LSTM with regularisation (dropout and max-in norm). The idea of max-in norm is to scale the weights of each unit in the network to have the maximum euclidean length of the dimensions of the input to these units and this is applied after each mini-batch. Other studies incorporate CNN [126], DBN [62], etc. Deep learning has shown high accuracy in HAR according to various studies. CNN is applied in [126] for the recognition of three basic movements collected from a single wrist-worn tri-axial accelerometer. The authors compare CNN with k-means, LDA and SVM and they shows that CNN has the highest performance amongst them. However, they did not clarify what features they used for k-means, LDA and SVM. Another CNN model is employed in [78] for online human activity recognition and they utilised statistical features as auxiliary input. They also prove that their model can work in real-time with small number of layers. The authors in [159] claim that simple deep neural network or MLP can outperform complex CNN since it can perform better feature extraction.

2.5 Machine Learning Classification Measures

The most commonly used measures fro classification performance and are useful for the problem discussed in this thesis is presented in this section. The measures include precision, recall, F1-score, receiver operating characteristic (ROC) curve [164], precision-recall (PR) curve [164] and area under curve (AUC) [164]. Precision [153], recall [153] and specificity [164] can be defined by the following equations 2.37 through 2.40:

$$Precision = pr(y = 1|\hat{y} = 1) \tag{2.37}$$

$$Recall = Sensitivity = pr(\hat{y} = 1|y = 1) \tag{2.38}$$

$$Specificity = pr(\hat{y} = 0|y = 0) \tag{2.39}$$

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall} \tag{2.40}$$

where, y is the true class label and $(\hat{y}$ is the predicted class label. The important point here is that recall and specificity are probabilities that are conditioned on the true class label to be positive in recall and negative in specificity. Whereas, Precision is a probability that is conditioned on the estimate of the predicted class label given that the true class label to positive. Hence, this will vary especially when predicting the classes in different populations with different baseline.

ROC curve [164] is a 2-dimensional plot that demonstrate the relation between the true positive rate (TPR, recall or usually called sensitivity) and the false positive rate (FPR). The FPR is on the x-axis and TPR is on y-axis. FPR is given by Equation 2.41. Precision-recall curve has the same concept as ROC curve in terms of it is 2-dimensional plot but demonstrate the relation between recall and precision. Recall is on the x-axis and precision is on the y-axis.

$$FPR = \frac{FP}{FP + TN} \tag{2.41}$$

, where TN is true negative and FP is false positive.

Chapter 3

Speech Detection in Naturalistic Environments

3.1 Introduction

3.1.1 Background and Motivation

Speech activity detection (SAD), also referred to as Voice activity detection (VAD) has recently received a great deal of attention and is a key component for many speech processing applications [99]. This line of research has received a fair amount of attention in recent years, particularly using deep learning approaches [44]. Furthermore, most standard speech detection methods show limited performance when applied to data recorded in naturalistic environments (i.e. outside of a clean recording studio) [58]. This limitation prevents their use in applications that require data to be recorded in unobtrusive environments such as, in the biomedical context, the analysis of the social interactions of individuals with Depression. Speech activity detection is used to identify speech segment boundaries from observed audio data. Most speech processing applications rely on speech detection as the first preprocessing stage such as speech coding, automatic speech recognition (ASR) and speaker verification [54]. SAD can be modelled as a classification task: assigning audio frames to the speech and non-speech categories.

Depression is associated with atypical language patterns, such as increased first-person singular pronoun use, more single-clause sentences, reduced utterances and incomplete phrases [162]. Patients with depression also show abnormal speech, including quieter speech, reduced variation of volume and pitch, and reduced prosody. Listeners who were naive to the depressive state of a speaker can perceive the severity of depression from vocal recordings of people with depression [183]. Clinical changes in depressive symptoms are

associated with differences in speech patterns and depression-related speech features can be found across different languages [191]. Acoustic speech analysis has been used to predict depression in at-risk participants two years before diagnosis with up to 74% accuracy [82] and automated analysis of language features can differentiate patients with schizophrenia and bipolar disorder from controls with 96% accuracy [172]. Most of the studies to date measure speech in controlled settings (e.g. in a quiet room during clinical interviews) and focus on detecting specific features of speech. An alternative approach would be to use wearable devices to objectively quantify how much speech participants encounter, and participate in, in their natural environment. Detecting speech this way could serve as a proxy for social interaction, encompassing numerous factors of social functioning that are often independently measured with different self-report scales.

3.1.2 Study Device

The private datasets used in this thesis are generated by an existing wearable sensing platform designed in Newcastle University and entitled the wearable activity monitor (WAM) shown in Figure 3.1. WAM includes a microphones for analysis of vocalised social interactions and 3D accelerometer for physical activity. The platform includes on-board memory that enables continuous recording over a whole week. As such the wrist-worn sensing platform has the potential to provide a more realistic and objective impression of everyday life activities (as a proxy for social interactions) of patients with psychological problems, which will enable improved diagnosis and treatment.

3.1.3 Approach

The conventional pipeline for speech detection starts with the extraction of relevant features followed by the modelling stage, typically using machine learning methods for distinguishing of speech/non-speech and finally the evaluation. There are some obstacles that hinder the process of developing robust speech detection methods as conventional feature engineering is unable to capture the changes in the audio environments particularly when it comes to dealing with unconstrained naturalistic-setting environments. The classification methods/representations should be sensitive to environment-produced distortions and have a highly efficacy to consolidate the temporal and spectral correlations on both the short and long-term levels[158].

In this chapter, a deep learning-based SAD method designed to tackle the challenging circumstances of naturalistic acoustic environments is presented. For that purpose, an end-to-end neural speech activity detection model for a robust speech detection. This model is

Fig. 3.1 WAM device.

called NatSpeech and designed for elderly depressed people in a conversation in naturalistic environments. The model is based on bidirectional LSTM with two different gated activations followed by LSTM layers mixed with maxout layers and Leaky ReLU layers. The input to the network is the normalised raw frames of speech data without any explicit feature extraction/engineering. The evaluation begins with assessment of standard SAD datasets from the literature captured in *clean* environments with/without artificial noise added on top. The evaluation includes a thorough comparison of the method against a range of SAD methods from the literature (from classic speech analysis to conventional machine learning and finally also deep learning methods). This is followed by evaluations of datasets that have been generated by the WAM device. Two datasets have been recorded using this device: one with volunteers that consented to humans listening to their naturalistic conversations (labelled the *discussion* dataset), and hence could be manually annotated with speech/non-speech class labels and one dataset of depressed patients from a clinical study run at Newcastle University and for which only automated processing is allowed to take place, without any human being able to listen to the audio and hence is un-annotated (labelled the *depression* dataset). I first evaluate the capacity of the method to successfully predict speech on the discussion dataset. Thereafter, the proposed SAD method can be effectively trained on the discussion dataset and then used to annotated the depression dataset in valuable ways. This will be presented in Section 3.4.

3.2 Method

3.2.1 Preprocessing

The WAM device stores the .EWV files and .OMX files on its storage. The .EWV format is the scrambled version (due to the ethical constraints) of the .WAV audio files which are 8KHz mono audio format. The .EWV files are descrambled first on the fly to create the small .WAV frames. These frames are 32ms length (i.e. 256 samples where the sampling rate is 8kHz=8192samples). The speech samples are uniformly rescaled in the range of [-1,1]. Subsequently, the frames are normalised according to the normal distribution with zero mean and unit variance over the whole training set. Thereafter, the frames become ready to be fed to the deep learning model.

3.2.2 Bi-Directional Recurrent Neural Networks (BRNN)

As discussed in Chapter 2 RNNs take time into consideration, bi-directional recurrent neural networks (BRNNs) also do the same but in two directions. The fundamental notation of BRNNs is to feed each training time step to forward recurrent networks and backward recurrent network. Both of these networks are connected to a third network or merely to the same output layer [144]. Each hidden layer of BRNNs has forward and backward layers. This leads to the fact that BRNN are able not only to leverage past history of the time steps in a given sequence but also to exploit the future context. Furthermore, finding such a task dependent time-window or target size delay is not even necessary due to the point that the network can use either as much or as little of the context as possible [59]. Deep BRNNs can be interpreted as replacing each hidden layer equation with the forward and backward layers as presented by the following equation:

$$y_\tau^{(l)} = [\overleftarrow{y_\tau^{(l)}}; \overrightarrow{y_\tau^{(l)}}] \tag{3.1}$$

where $\overleftarrow{y_\tau^{(l)}}$ is the forward layer and $\overrightarrow{y_\tau^{(l)}}$ is the backward layer.

3.2.3 Gated Bi-Directional LSTM (GBLSTM)

Bi-Directional LSTMs (BLSTM) is a combination of BRNNs with LSTMs. That means LSTM blocks are used for the hidden layer of the BRNNs as shown in Fig. 3.2. The main idea of GBLSTM is that the sequences are fed to the gated activation unit (see Fig. 3.3). That is, the sequences are fed to two bi-directional LSTM layers, one with hyperbolic

tangent activation and the second with logistic activation. The two layers are then normalised by l_2 regularisation. Following that is the merge layer which is another gated activation unit. The gated activation unit is the the element-wise product of hyperbolic tangent and logistic/sigmoid activations. The gated part is utilised in CNN to substitute the output gate of LSTM and outperforms LSTM [34]. The rationale of gated part here with BLSTM is to allow the network to control what data to be let through the hierarchy of layers in addition to the output gate in LSTM. This enables the network to control what features are relevant to predict the speech frame based on its past and future frames. So, this works as an extra output gate for BLSTM. Besides that, it helps to discriminate between the noisy speech frame followed by another speech frame or followed by audible speech which is considered as non-speech in this research. The use of Maxout [55] here improves the optimisation of dropout and helps for the final classification layer.

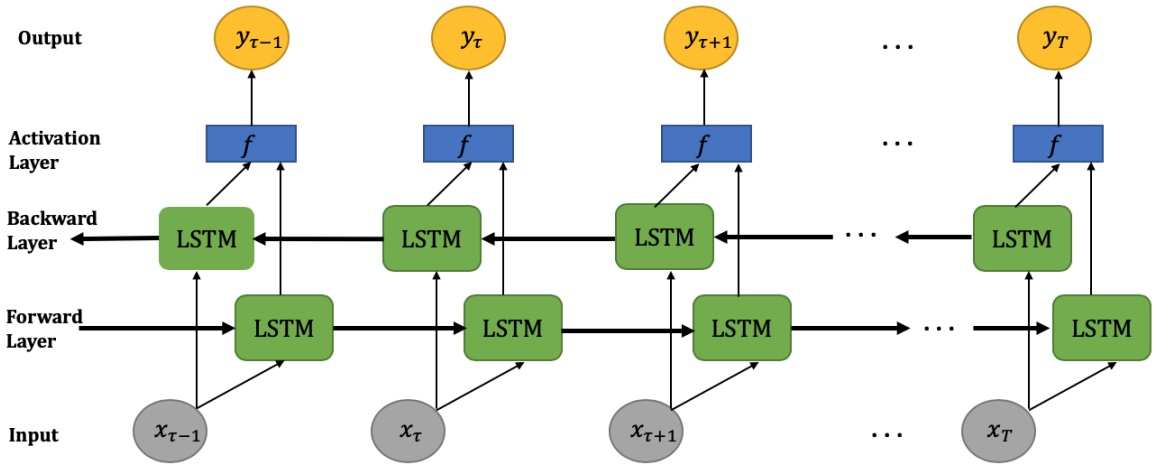

Fig. 3.2 Bidirectional Long Short-Term Memory Network (BLSTM). The forward pass is the same as for a unidirectional LSTM, except that the input sequence is presented in opposite direction for the backward pass and then fed to the two hidden layers. The output layer is not updated until both hidden layers have processed the entire input sequence.

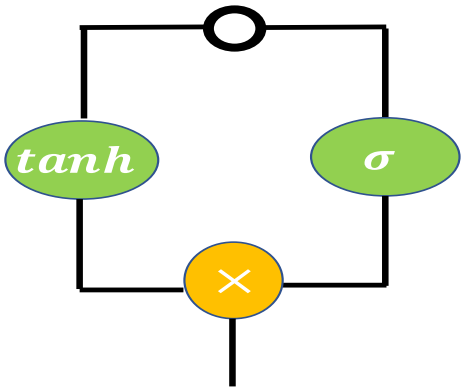

Fig. 3.3 Gated Activation Unit. The circle in the figure represents the input flows to the the two layers one with hyperbolic tangent activation and the other with sigmoid activation. The output from these two layers is merged with element-wise product process.

3.2.4 Transfer Learning

Transfer learning is a research area in ML that transfer the knowledge gained from a model trained and validated on specific data for specific task to be applied to different data or another related task [100, 54]. This means that the pretrained model is reused as a starting point especially for computation vision or natural language processing tasks and allows for improved performance [53]. In transfer learning, the base model is trained on the base task with base dataset. The learned features from the base model is then transferred to the target model on a target dataset [186]. Transfer learning is used in this thesis to create a pretrained model on the discussion dataset to be be reapplied to the depression dataset due to the ethical constraints. First, the proposed model is trained and validated on the discussion dataset. Thereafter, the pretrained model with the same architecture and its parameters is then applied to the depression dataset as discussed in Section 3.4.

3.2.5 Network Architecture

The architecture starts with initial experiments on 6 layers of LSTM, then the Dropout and gradient normalisation have been added to them to get better F1-score. Then Maxout layers are added with Dropout layer by layer until there is no improvement in F1-score. Then one fully connected layer has been added with different activation until it there is improvement occurred with leaky ReLU activation function. For this layer, batch normalisation is decided using the visualisation tool in Deeplearning4j. Then the bidirectional LSTM layers are added one by one as before with the same dropout and gradient normalisation like in LSTM. By looking at the predictions of the classifier which stuck at the prediction of audible speech

within background as non-speech which is challenging case. The reason to exclude them is that the main focus is the speech of the wearer and people to them not the other people who are not included in the conversation. This is solved by the gated activation layer which works as an extra output gate for LSTM. The number of gated layers is decided based on F1-score through the validation. The final architecture is displayed in Fig. 3.4 and the rationale for each part is discussed above.

During the training process, the preprocessed frames are fed to a TANH layer and then to a SIGMOID layer of the gated activation unit block. Thereafter, these layers are regularised using L_2 regularisation. Then the output of the gated unit is the element-wise product of the activation as stated in the previous section. The gated activation is used twice as illustrated in Fig. 3.4. This novel part of the structure provides a large improvement on the prediction performance of the network as will be shown in the experimental evaluation. The usage of bi-directional LSTM might not have the big difference in the performance from using the uni-directional as the future samples of the speech frames will not have the same influence like the past history. However, they assist in the cases of very noisy data and potentially the audible speech in the background which represents a great challenge too.

The output of the merge layer is fed to four layers of bi-directional LSTM. Dropout is exploited with gradient normalisation. Following these layers, there are 6 layers of unidirectional LSTM with dropout and gradient normalisation are used. The fully connected layers are also used in our model but the maxout activation was leveraged in the next three layers with dropout followed by a perceptron layer with leaky RELU activation. The final layer is a softmax one. In this model cost sensitive classification is leveraged where weighted loss is used. Overall NatSpeech architecture is represented in figure 3.4. The source code of our method is available at https://github.com/OssamaAlshabrawy/NatSpeech.

3.2.6 Network Training

The training phase is performed by backpropagation through time (BPTT) and the loss function used is cross-entropy. Adam optimisation algorithm is utilised in the layers from the first hidden layer until the last layer of unidirectional LSTM and the learning rate is 10^{-5}. The learning rate is chosen based on grid search with the values $(10^{-1}, 10^{-2}, ..., 10^{-6})$. The final four layers (maxout and dense) are trained using RMSPROP and the learning rate is 10^{-4}. Over-fitting of the training set is managed by early stopping algorithm where we choose the maximum number of epochs to be 100. For parameter regularisation, L_2 regularisation and dropout with 0.7 retain probability. The dropout is chosen using a grid search with the retaining values (0.2, 0.3, 0.4, 0.5, 0.6, 0.7 and 0.8). In addition, batch normalisation which normalise the activation of the previous layer at each batch (covariance

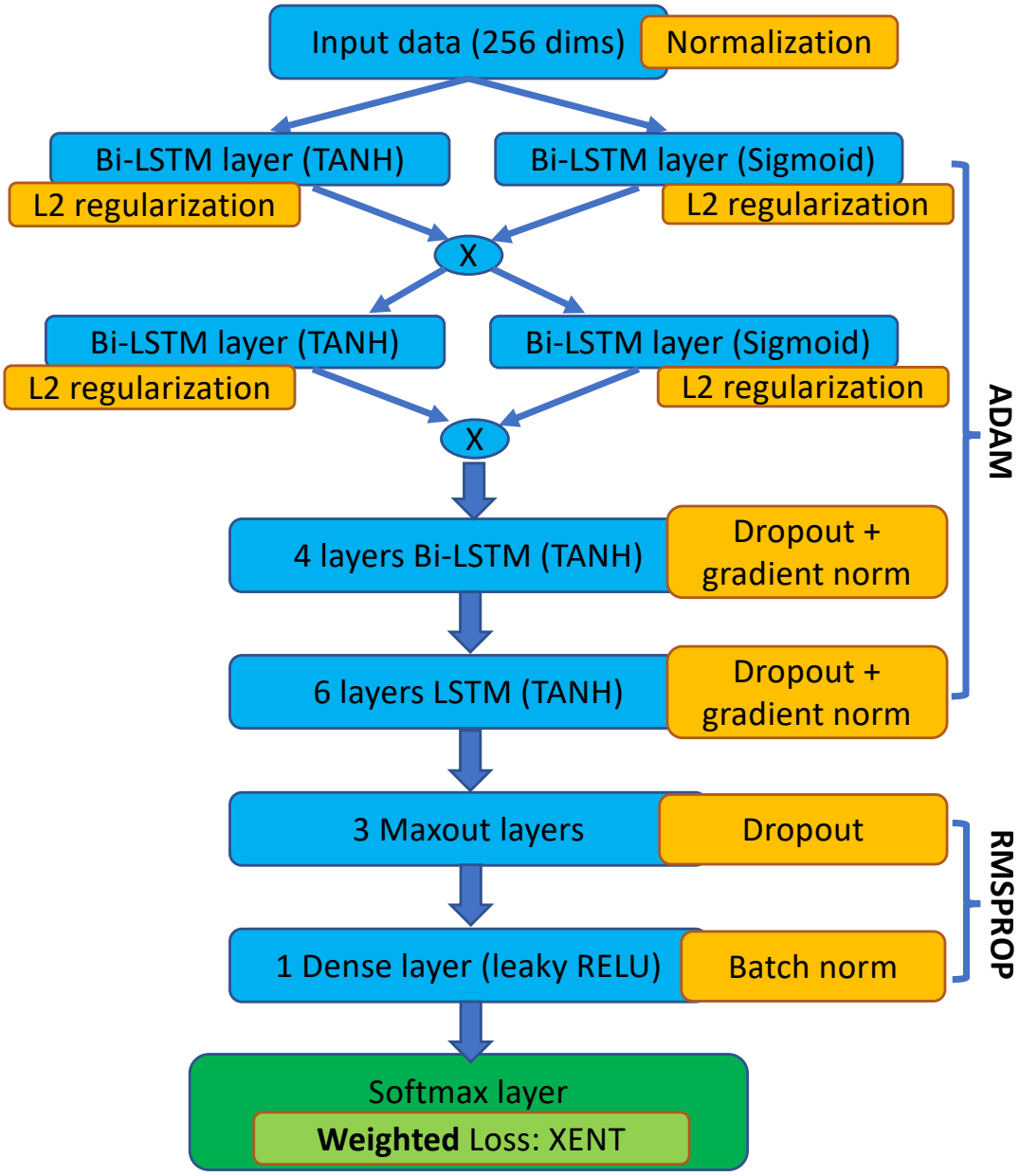

Fig. 3.4 Deep Learning Model Architecture.

shift i.e. transform the activations to have a mean close to 0 and standard deviation close to 1), is utilised in order to speed up the training. Likewise, gradient normalisation is exploited which is applied to the raw gradients before passing them to the optimisation algorithm. The gradient normalisation is utilised here by dividing the gradient by L_2 of all gradient for the layer. The number of LSTM cells in each layer is a number between 128 and 256. The framework used for implementing the architecture is Deeplearning4j [43]. The number of LSTM cells is optimised by HyperParameterOptimization tool in Deeplearning4j. The loss function is weighted with 0.6 for non-speech and 0.4 for speech as in most of the sessions of the dataset, we have more speech frames than non-speech ones.

3.3 Datasets

3.3.1 Main Datasets

The main dataset of interest (namely, the *depression dataset*) has been recorded in a previous project where 30 participants with diagnosed clinical depression, and 30 healthy participants (healthy control group) wore the WAM device resulting in a dataset of audio data. In addition, for this set of patients there is accelerometer data, which will be discussed later in Chapter 5. A whole week of naturalistic audio was recorded for each participant. As for privacy reasons, the data is stored in a scrambled format to only allow ethically approved automated audio processing. Hence, it is impossible to annotate the dataset with class labels (e.g. speech/non-speech, or whether the person talking is the wearer/non-wearer).

In order to overcome this limitation imposed by the ethical constraints and be able to predict the speech patterns of this dataset, we need to train our proposed deep learning model on similar data, ideally recorded by the same device and also in the same naturalistic settings, that is not constrained by the same ethical constraints. One solution is to use the WAM device without activated speech scrambling and record a second dataset where volunteers have given full consent allowing for systematic system development and validation. The dataset we have created (namely the *discussion dataset*) contains 15 sessions of non-sensitive, staged, discussions with a total length of 20 hours of recording, created using the same settings as the depression dataset stated above. The sessions details are shown in the following two tables (Table 3.1 and Table 3.2.

Table 3.1 Discussion Dataset recording environment and percentage of total speech and non-speech respectively.

Session	Circumstances	speech (%)	non-speech (%)
1	Indoor session (eating conversation in the common kitchen)	52	48
2	Indoor session (with audible speech within background)	28	72
3	Indoor session (with audible speech within background)	20	80
4	Outdoor session (walking around from the university and through a shopping centre)	63	37
5	Outdoor session (conversation in a car with radio on and windows open for some time, then walk around)	47	53
6	Outdoor session (walking around with different types of noises like cars)	55	45
7	Indoor session (noisy room)	85	15
8	Indoor session (noisy room with audible speech within background)	90	10
9	Outdoor session (walking down the city centre)	79	21
10	Indoor session (noisy room)	82	18
11	Indoor session (noisy room)	63	37
12	Outdoor session (with clear audible speech within background)	28	71
13	Outdoor session (walking down the city centre)	88	12
14	Indoor session (a little bit of noise)	70	30
15	Indoor session (a little bit of noise)	13	87

Table 3.2 Discussion Dataset recording environment and percentage of speech produced by the wearer or no-wearer respectively.

Session	Circumstances	Percentage of speech produced by wearer & non-wearer respectively (%)	
1	Indoor session (eating conversation in the common kitchen)	45	55
2	Indoor session (with audible speech within background)	45	55
3	Indoor session (with audible speech within background)	36	64
4	Outdoor session (walking around from the university and through a shopping centre)	40	60
5	Outdoor session (conversation in a car with radio on and windows open for some time, then walk around)	30	70
6	Outdoor session (walking around with different types of noises like cars)	52	48
7	Indoor session (noisy room)	31	69
8	Indoor session (noisy room with audible speech within background)	22	78
9	Outdoor session (walking down the city centre)	10	90
10	Indoor session (noisy room)	43	57
11	Indoor session (noisy room)	45	55
12	Outdoor session (with clear audible speech within background)	62	38
13	Outdoor session (walking down the city centre)	28	72
14	Indoor session (a little bit of noise)	43	57
15	Indoor session (a little bit of noise)	41	59

3.3.2 Public datasets for extra evaluation

The public datasets used in this chapter to evaluate the proposed method are Aurora 2 [128], Aurora 4 [129] and TIMIT [148]. Aurora 2 contains audio samples of the pronunciations

of digits while Aurora 4 contains the utterances of continuous speech. Several variants of Aurora 2 with varying degree of noise are used, at different signal to noise ratio (SNR) levels (-5, 0, 5, 10) dB. The sampling rate is 8kHz. The data is split into 300, 300 and 401 utterances for training, validation and test sets, respectively. The Aurora 2 noisy data is established by randomly choosing and concatenating 4000 noise segments to form a long noise waveform of roughly 35 hours; then the noisy segments are artificially added to the clean utterances of Aurora 2 clean corpus under the different SNR levels stated before. The noise segments are generated from NOISEX-92 dataset [155].

Aurora 4 data is corrupted by the factory and babble noises from the NOISEX-92 dataset at the same SNR levels stated before in order to have challenging and broaden comparison with the state-of -art methods. The sampling rate for these data is 16 kHz. The clean speech corpus comprises 7,138 training utterances and 330 test utterances. The Aurora 4 is constructed with the same procedure described for Aurora 2 above. The ground-truth labels for both Aurora 2 and Aurora 4 are generated by the Sohn [152] VAD method following the evaluation protocol in [187] and [85].

The TIMIT corpus contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States. The dialect regions are New England, Northern, North Midland, South Midland, Southern, New York City, Western and Army Brat. The distribution of the speakers is 70% and 30% for male and female respectively. The ground-truth labels are generated by the Sohn [152] VAD method following the evaluation protocol in [187]. As for training and testing purposes, the data is partitioned into 3696 utterances for training and 1344 for testing. The sampling rate for the utterances is 16kHz. The training data is split into training and validation based on the dialect regions. In our experiments, dialect region 1 (New England region) is used as a validation data and the rest for training as suggested in [86]. The test set has a core portion containing 24 speakers, 2 male and 1 female from each dialect region. This set is totally excluded from the training and validation processes. The source of noise used in this dataset is the NOISEX-92 dataset similarly to Aurora 4. These noises include white noise, babble and Volvo. The noise is added to the speech data using the FaNT tool [71]. The frame setting for all of these public datasets is 25ms window with 10ms overlapping (frame shift).

3.4 Results and Discussions

The evaluation of the proposed method is performed to the public datasets followed by evaluation on the main datasets: the discussion dataset then the Depression dataset. These

datasets have been recorded in naturalistic environments and present a variety of background sounds, even speech (e.g. the sound of TV in the background).

3.4.1 Results on the public datasets

The first set of experiments is conducted on Aurora 2, Aurora 4 and TIMIT datasets. The reason is to gauge first the application of the proposed system on standard datasets. Table 3.3 shows the F1-score on test data of the proposed method in comparison with different SAD methods on the 42 noisy environments of AURORA2. Table 3.4 shows the F1-score result but this time on the 8 noisy environments of AURORA4. To help summarise the performance of all methods included in these experiments we added two rows that the report (1) the average rank (lower rank is better) of each method across datasets and (2) the number of times that each method obtained the top performance. Our proposed model (GBLSTM) has the top rank in these experiments, followed by the two most recent deep-learning based methods from the literature (Sehgal18 and Gelly18). The proposed method by far is the one obtaining more times the top performance across datasets.

The performance of our SAD method is compared against a variety of methods that range from classic methods such as Sohn99 [152], Random Forest (RF), Support vector Machines (SVM), Ramirez05 [133] as well as more recent methods such as Ying11 [185], Multi-resolution stacking (MRS) [187], Sehgal18 [146], Gelly18 [52], the latter three based on deep learning. The parameter settings for SMV are (regularization parameter:0.8, kernel:radial basis function, gamma:1/256). The parameter settings for RF are (number of trees: 200, criterion: Gini impurity, max depth of the trees: 6, maximum number of features: 25). The parameter settings for Ramirez and Ying are the same as in [187]. For MRS-based VAD, the number of building blocks: 2, For the bottom one, the number of boosted deep neural networks is 10, the resolution parameter is set to [(3,1), (5,2), (9,4), (13,6), (15,7), (17,8), (19,9), (21,10), (23,11), (25,12)] respectively. The numbers of hidden units:512, the number of epochs: 50, batch size: 512, scaling factor for the adaptive stochastic gradient descent: 0.0015, and the learning rate decreased linearly from 0.08 to 0.001. The momentum of the first 5 epochs was set to 0.5, and the momentum of other epochs was set to 0.9. The dropout rate of the hidden units was set to 0.2.

MRS is an ensemble of LSTM classifiers are stacked in which each of these classifiers take the soft output predictions and concatenate them with the expansion of the acoustic features. Sehgal et al. (referred to as Sehgal18 in the results) [146] proposed a CNN architecture that takes the audio mel-energy spectrum images and pass it to the network which comprises 3 convolution layers, pooling layer and one fully-connected and 2 softmax layers. The idea is relying on using Fourier transform and then obtaining the spectrogram

images and then deal with the audio as imaging data to be fed to two dimensional CNN layers. Gelly et al. [52] utilise quantum-behaved particle swarm optimisation is utilised for optimising the feature extraction. That is followed by the initial training of the network and then pass to two different models of LSTMs. However, in GBLSTM, the raw data is fed directly to the network so, the architecture proposed requires minimal preprocessing compared to Sehgal18 and Gelly18. Compared to MRS which requires feature extraction, the idea of GBLSTM does not utilise ensemble LSTM classifiers which requires more training time. Moreover, I evaluate several variants of our method to assess the individual contribution of (1) using BLSTM layers rather than unidirectional LSTMs and (2) gated layers rather than standard ones. In total four variants of our method are tested, two non-gated (labelled as LSTM and BLSTM) and two gated (labelled as GLSTM and GBLSTM).

Table 3.3 Performance evaluation (F1) on Aurora 2 dataset. The best method for each dataset is marked in bold. Last row based on a Holm post-hoc statistical test at 95% confidence level

Noise	SNR	**SVM**	**RF**	**Sohn99**	**Ramirez05**	**Ying11**	**MRS16**	**Sehgal18**	**Gelly18**	**LSTM**	**BLSTM**	**GLSTM**	**GBLSTM**
Babble	-5 dB	67.03	69.44	60.45	61.33	59.17	77.92	84.46	85.98	75.44	85.34	85.59	**86.45**
	0	74.23	76.56	68.66	70.08	65.63	88.66	90.65	**91.77**	88.69	90.45	89.87	91.16
	5	84.33	85.98	79.83	81.94	76.52	92.34	93.21	93.87	91.5	93.13	93.06	**94.77**
	10	86.76	87.88	86.76	88.12	84.46	94.65	96.98	96.04	94.56	96.37	96.45	**97.03**
Car	-5	78.65	79.48	59.03	60.62	61.75	88.55	89.47	89.76	88.78	89.64	89.86	**91.12**
	0	85.53	86.98	69.05	72	69.27	94.06	95.36	95.17	94.23	94.56	93.67	**96.23**
	5	88.54	90.62	79.83	82.22	78.53	95.66	96.23	96.34	94.89	95.98	96.23	**96.88**
	10	90.98	92.08	87.22	88.64	84.73	96.34	97.23	**97.98**	95.97	97.56	97.24	97.89
Restau-rant	-5	68.45	70.76	55.62	55.04	57.67	73.06	80.73	82.87	75.78	81.89	82.43	**85.45**
	0	76.22	77.97	66	64.52	62.98	84.67	87.03	88.65	85.17	88.67	87.84	**89.33**
	5	85.52	88.31	72.24	73.65	71.54	90.86	91.56	**93.12**	90.77	91.21	91.45	93.08
	10	87.33	90.14	79.85	80.85	79.29	93.12	95.34	94.98	92.67	94.76	95.45	**96.89**
Street	-5	68.76	71.76	53.68	54.8	55.63	79.89	82.56	83.44	78.75	81.93	80.65	**85.76**
	0	76.41	78.12	60.03	60.06	61.62	88.16	90.35	91.54	90.42	89.65	89.76	**92.17**
	5	85.13	87.09	68.74	71.52	70.28	91.34	92.45	92.23	90.89	91.76	91.23	**94.67**
	10	87.09	89.87	76.04	78.21	76.41	93.66	**97.02**	96.45	94.06	95.86	95.34	97.01
Airport	-5	70.01	72.43	56.6	59.93	59.06	80.22	83.58	84.23	80.17	85.43	85.67	**86.66**
	0	77.13	79.65	64.22	66.11	66.07	87.44	88.02	88.87	88.01	88.96	88.53	**90.13**
	5	86.04	86.01	73.78	76.9	74.48	91.67	92.12	92.95	90.43	92.07	91.89	**93.98**
	10	87.54	86.54	83.18	86.06	84.21	94.33	95.68	**96.23**	94.79	95.21	94.23	96.08
Train	-5	70.89	71.98	55.31	57.68	61.35	81.67	82.45	83.03	81.87	83.78	82.67	**84.87**
	0	78.35	80.53	60.04	63.19	67.85	88.65	89.11	90.07	85.76	89.64	90.11	**91.31**
	5	86.79	85.98	73	77.26	77.58	91.78	92.53	93.74	89.76	92.76	92.87	**94.34**
	10	88.16	90.41	83.76	84.51	82.18	96.89	94.45	95.98	93.46	95.78	95.89	**96.87**
Subway	-5	71.38	73.87	55.42	55	57.74	75.98	80.54	81.21	76.87	82.56	82.67	**85.23**
	0	79.14	83.54	62.66	61.63	62.75	84.31	87.98	86.32	85.88	86.58	86.34	**89.53**
	5	87.28	85.98	70.49	76.5	68.35	89.82	90.67	91.77	90.64	90.93	90.96	**92.01**
	10	89.89	93.65	79.02	81.18	76.73	93.63	96.03	**96.43**	94.12	95.24	95.36	95.23
Avg Rank		8.91	8.14	11.41	10.39	11.11	6.25	3.77	2.57	6.32	3.86	3.91	**1.36**
Number of times best		0	0	0	0	0	1	1	5	0	0	0	21
always beaten by GBLSTM		yes	yes	yes	yes	yes	yes	yes	no	yes	yes	yes	-

Table 3.4 Performance evaluation (F1) on the Aurora 4 dataset. The best method for each dataset is marked in bold. Last row based on a Holm post-hoc statistical test at 95% confidence level

Noise	SNR	**SVM**	**RF**	**Sohn99**	**Ramirez05**	**Ying11**	**MRS16**	**Sehgal18**	**Gelly18**	**LSTM**	**BLSTM**	**GLSTM**	**GBLSTM**
Babble	-5 dB	77.84	80.76	70.69	75.9	64.63	82.67	84.65	84.34	82.23	84.78	83.76	**86.97**
	0	82.77	85.97	77.67	83.05	70.72	85.65	87.67	87.21	85.87	86.59	86.78	**90.02**
	5	86.04	86.01	84.53	87.85	78.7	88.87	89.98	90.33	88.21	90.31	89.93	**93.34**
	10	88.01	90.54	89.18	89.93	85.61	91.88	**96.76**	95.87	93.74	95.89	95.76	96.73
Factory	-5	74.43	73.77	58.17	58.37	62.56	81.74	84.77	84.65	82.43	83.98	85.12	**85.93**
	0	82.89	86.13	64.56	67.21	68.79	85.01	87.65	88.54	86.27	87.88	87.89	**89.85**
	5	86.04	85.97	72.92	76.82	75.83	88.9	91.78	**92.79**	89.46	90.67	88.69	92.65
	10	88.19	91.65	80.8	84.72	82.64	91.61	92.27	93.5	91.63	91.76	91.35	**94.49**
Avg Rank		9.12	7.75	11.38	9.75	11.25	6.88	3.00	2.75	6.25	3.75	4.88	**1.25**
Number of times best		0	0	0	0	0	0	1	1	0	0	0	6
always beaten by GBLSTM		yes	yes	yes	yes	yes	yes	no	no	yes	no	no	-

From the results on Aurora 2, it is clear that Gelly18 outperforms in the cases of high SNR which means if the speech is cleaner and cannot outperforms GBLSTM in cases of noisy speech or low-SNR. Although Gelly18 outperforms only 5 times and GBLSTM 21 times, however, by applying Holm post-hoc test it proves that GBLSTM is not always beating Gelly18. In Aurora 4, GBLSTM is not always beaten by 4 methods but the reason is due to the small number of experiments in this dataset compared to Aurora 2. As shown in the Table 3.4 Sehgal only outperforms once for babble noise and in high SNR and the same for Gelly18 which is also outperforms in case of factory noise. These methods outperforms in case of clean speech which is not the case of the main objective of this thesis where this thesis focuses more on working out the free-living or naturalistic data with different noises. Likewise, these methods don't take into account excluding the audible speech in the background as in GBLSTM.

Table 3.5 Performance evaluation (F1) on the TIMIT dataset. The best method for each dataset is marked in bold. Last row based on a Holm post-hoc statistical test at 95% confidence level

Noise	SNR	**Sehgal18**	**Gelly18**	**LSTM**	**BLSTM**	**GLSTM**	**GBLSTM**
White	15	93.54	92.65	89.76	90.54	89.65	**97.87**
	10	92.57	**94.62**	87.68	93.65	95.78	94.17
	5	91.89	90.56	88.87	92.14	90.57	**93.78**
	0	86.33	91.67	86.34	85.36	85.79	**93.45**
Volvo	15	94.05	94.11	90.34	90.23	90.56	**97.76**
	10	92.45	92.87	89.78	89.45	87.28	**96.88**
	5	91.66	**92.01**	89.43	90.31	89.98	91.76
	0	88.53	91.47	85.48	87.59	88.63	**93.43**
Babble	15	93.31	93.54	89.32	90.65	91.35	**96.45**
	10	92.87	92.76	90.37	89.54	89.76	**94.67**
	5	91.98	92.02	89.39	90.45	88.67	**93.76**
	0	90.87	**91.65**	85.31	85.17	85.21	91.56
Avg Rank		3.17	2.25	5.00	4.75	4.50	1.33
Number of times best		0	3	0	0	1	9
always beaten by GBLSTM		yes	no	yes	yes	yes	-

The second evaluation for the proposed model is achieved on TIMIT dataset. For this dataset, I restricted the comparison to recent deep learning methods from the literature as they dominated the ranking of performance in Aurora 2 and 4 datasets. The results reported for the TIMIT datasets in table 3.5 are consistent with the previous result. The proposed method (GBLSTM) variant obtained the top performance followed by Gelly18 and Sehgal18. Given that the partitioning of this data into training and test sets was done by dialect, these results highlight the robustness of our method to predict speech across a variety of scenarios (dialects, noise). Also, from the table only Gelly18 outperforms GBLSTM for 3 times but this dataset doesn't have lower SNR like in Aurora 2 and Aurora 4.

To assess the statistical significance of our method's performance we run the Friedman statistical test for multiple comparisons, following the recommendations in [36]. This test it designed to assess the performance of multiple methods over multiple datasets, and assesses whether there are significant differences in performance between methods. In all datasets significant differences were identified: Aurora 2 (p-value $= 2.2 \times 10^{-16}$), Aurora 4 (p-value $= 2 \times 10^{-12}$), and TIMIT (p-value $= 3 \times 10^{-07}$). A Holm post-hoc test was then applied to determine, with corrections for multiple comparisons, if the performance of the best method was significantly better than the performance of the other methods with 95% confidence. The last row in tables 3.3, 3.4 and 3.5 indicates the methods significantly outperformed by GBLSTM. In summary, GBLSTM was the best method in all datasets and significantly outperformed the of other methods, with the exception of Gelly18.

3.4.2 Results on Discussion Dataset

The nature of the sessions as well as the percentage of speech in the discussion datasets is presented in the datasets section in Table 3.1. The dataset shows a variety of situations that represent the real-world recording scenario. In the discussion dataset we employ a leave-one-session-out cross-validation methodology to obtain a robust estimate of speech prediction capacity in this dataset of naturalistic nature. That means the 14 sessions are used for training and there is one session left for validation/test. For each test session I compute accuracy, precision, recall, F1-score and area under the curve (AUC). Table 3.6 reports the detailed results of our GBLSTM method across all sessions.

Table 3.6 Performance evaluation measures of the GBLSTM method on the discussion dataset

Session	Accuracy	Precision	Recall	F1-Score
1	0.89	0.93	0.86	0.89
2	0.943	0.85	0.88	0.86
3	0.94	0.92	0.94	0.93
4	0.84	0.81	0.98	0.89
5	0.86	0.87	0.86	0.86
6	0.88	0.85	0.96	0.9
7	0.95	0.95	0.96	0.95
8	0.95	0.96	0.95	0.95
9	0.9	0.92	0.95	0.93
10	0.92	0.94	0.92	0.93
11	0.97	0.97	0.99	0.98
12	0.92	0.94	0.92	0.93
13	0.93	0.94	0.93	0.93
14	0.94	0.95	0.94	0.94
15	0.95	0.96	0.97	0.96

The plots of ROC-curves as well as Precision-Recall curves with AUC for some selected sessions in the dataset are shown in Figs. 3.5 to 3.9. The rest of the session is shown in the appendix Section A.1. Session 4 which shows a lower performance is due to the challenging case of recording where the data is recorded in a shopping street with a lot of audible speech in the background. That definitely represents a complex case because of the different noises mixed with the speech in addition to the audible speech which need to be excluded. Session 5 also shows less performance where it contains another challenging case of noisy speech in

a car with windows open and the radio is open for some time too while in session 12 there is more challenging where the audible speech is presented. On the other hand, the indoor session which has audible speech in a noisy room can be also considered as a challenging case for the model.

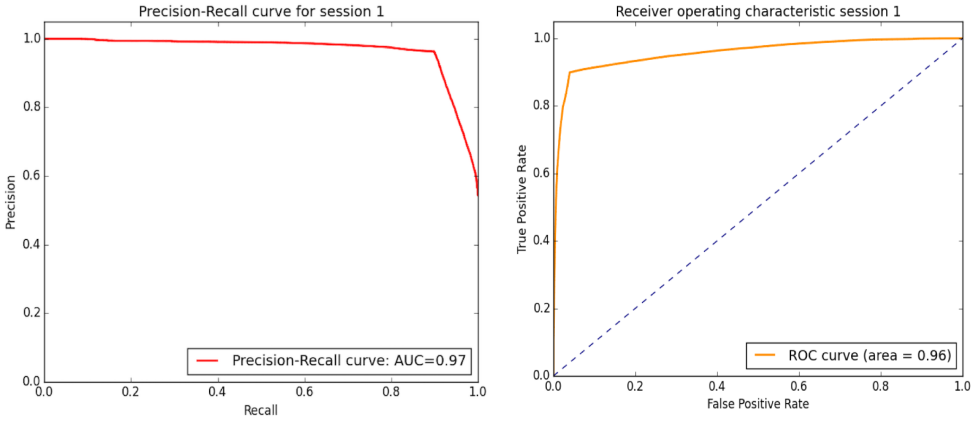

Fig. 3.5 ROC and PR curves for session 1 (Discussion Dataset)

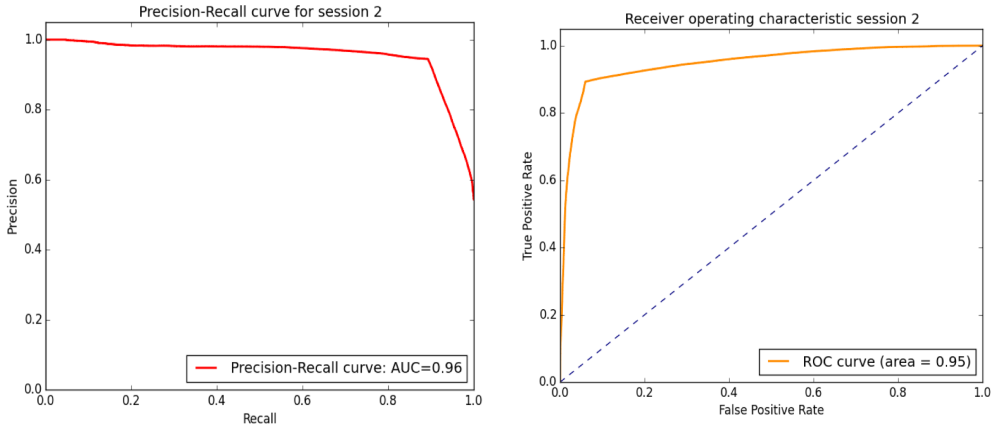

Fig. 3.6 ROC and PR curves for session 2 (Discussion Dataset)

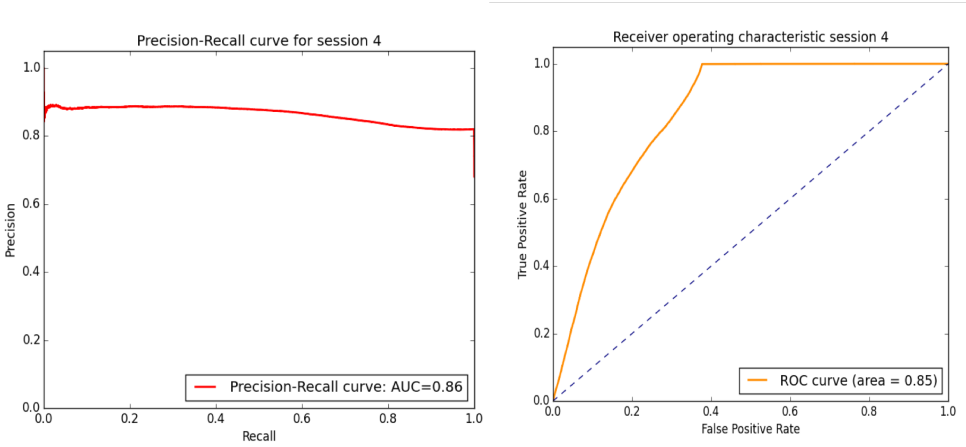

Fig. 3.7 ROC and PR curves for session 4 (Discussion Dataset)

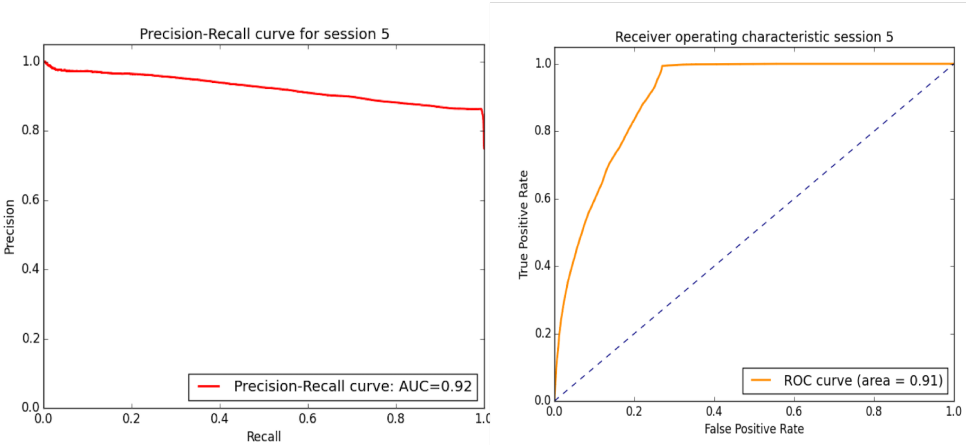

Fig. 3.8 ROC and PR curves for session 5 (Discussion Dataset)

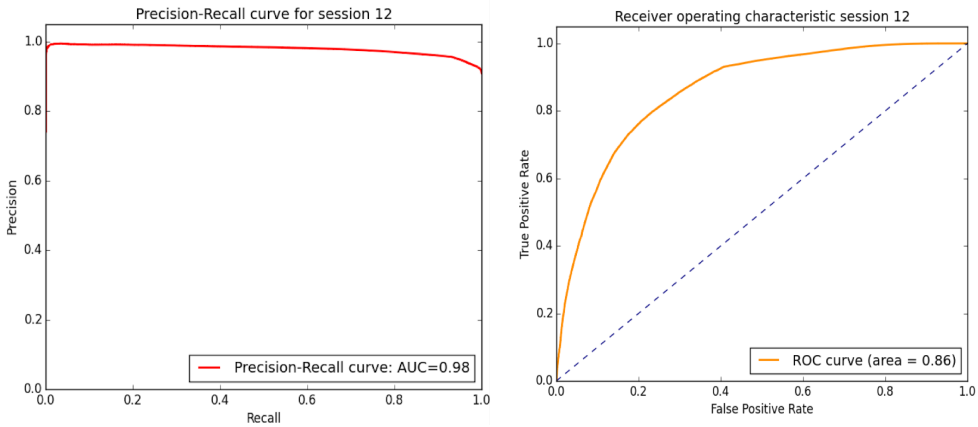

Fig. 3.9 ROC and PR curves for session 12 (Discussion Dataset)

Table 3.7 Performance evaluation (F1) on the Discussion dataset. The best method for each dataset is marked in bold. Last row based on a Holm post-hoc statistical test at 95% confidence level

Session	Sehgal 18	Gelly 18	LSTM	BLSTM	GLSTM	GBLSTM
1	0.84	**0.91**	0.82	0.87	0.85	0.89
2	0.83	**0.87**	0.84	0.85	0.84	0.86
3	**0.94**	0.89	0.87	0.91	0.9	0.93
4	0.87	0.87	0.86	0.85	0.88	**0.89**
5	**0.89**	0.85	0.84	0.86	0.87	0.86
6	0.89	0.89	0.86	0.88	0.88	**0.9**
7	0.9	0.91	0.88	0.89	0.9	**0.95**
8	0.89	**0.95**	0.84	0.9	0.91	**0.95**
9	**0.94**	0.91	0.88	0.87	0.9	0.93
10	0.84	0.86	0.82	0.92	**0.94**	0.93
11	0.94	0.93	0.9	0.94	0.93	**0.98**
12	0.9	0.89	0.89	**0.95**	0.94	0.93
13	0.93	0.94	0.9	0.93	0.95	**0.96**
14	0.93	0.91	0.89	0.92	0.95	**0.97**
15	0.92	0.93	0.87	0.91	0.9	**0.96**
Avg Rank	3.37	3.23	5.73	3.80	3.20	1.67
Number of times best	3	3	0	1	1	7
Always beaten by GBLSTM	yes	yes	yes	yes	yes	-

In addition to the ROC and PR curves for the proposed method, F1-score comparison between the proposed method and other recent literature and is shown in table 3.7. In this case I only compare the variants of our method to Sehgal18 and Gelly18 because they are the best methods in the previous experiments using public datasets. The results on the discussion dataset are largely consistent with the previous experiments. GBLSTM and GLSTM methods are the best, followed by Sehgal18 and Gelly18. By analysing the difference in performance between the variants of our architecture we can see that the gated component is the most important as both GBLSTM and GLSTM have better rank than LSTM and BLSTM. To a lesser extent the bi-directional units also contribute to better performance. I also performed a Friedman test on this dataset similar to the tests above, which again detected significant

differences across methods (p-value $= 4 \times 10^{-07}$). The Holm post-hoc test detected that GBLSTM significantly outperformed all other methods with 95% confidence.

3.4.3 Results on Depression Dataset

After the evaluation on the *discussion* and public datasets that shows high performance, the system is then applied to the participants data of the Depression dataset. Figure 3.10 represents the heatmap of the predictions of speech for the controls and depressed cohort for the whole day. This one day predictions is produced averaging the predictions of the seven days together. The numbers from 0 to 28 on Y-Axis represent the controls subjects and number in the range of 29 to 57 represents the depressed cohort. The order of the rows in the heatmaps is automatically defined by a hierarchical clustering algorithm, that generates a dendrograms also shown in the figure. The figure shows that the hierarchical clustering, working only on the speech predictions, is able to identify two subgroups of samples with perfect separation between the depressed patients and the healthy controls.

In order to give a complete overview of the Depression Dataset and the difference between the two groups (healthy controls and LLD patients), Table 3.8 displays group demographics(Sex, live alone, age, National Adult Reading Test (NART), education and Mini Mental State Exam (MMSE)), clinical characteristics, self-reported social functioning and speech data. The demographics and clinical and social characteristics statistical analysis of the table are regenerated from the work published in [121]. The two groups did not differ in sex, living status, handedness, age, or premorbid IQ. LLD patients group had fewer years of education and lower MMSE scores than controls. LLD scored higher than controls on UCLA Loneliness Scale (UCLA-LS), reflecting higher self-reported loneliness. LLD scored lower than controls on both depression scales: Montgomery-Asberg Depression Rating Scale (MADRS) and Geriatric Depression Scale (GDS-15), general health and functioning: Short-Form Health Survey (SF-36) and Instrumental Activities of Daily Living (IADL), and self-reported social interaction and social network: Duke Social Support Index (DSSI) and Lubben Social Network Scale-Revised (LSNS-R).

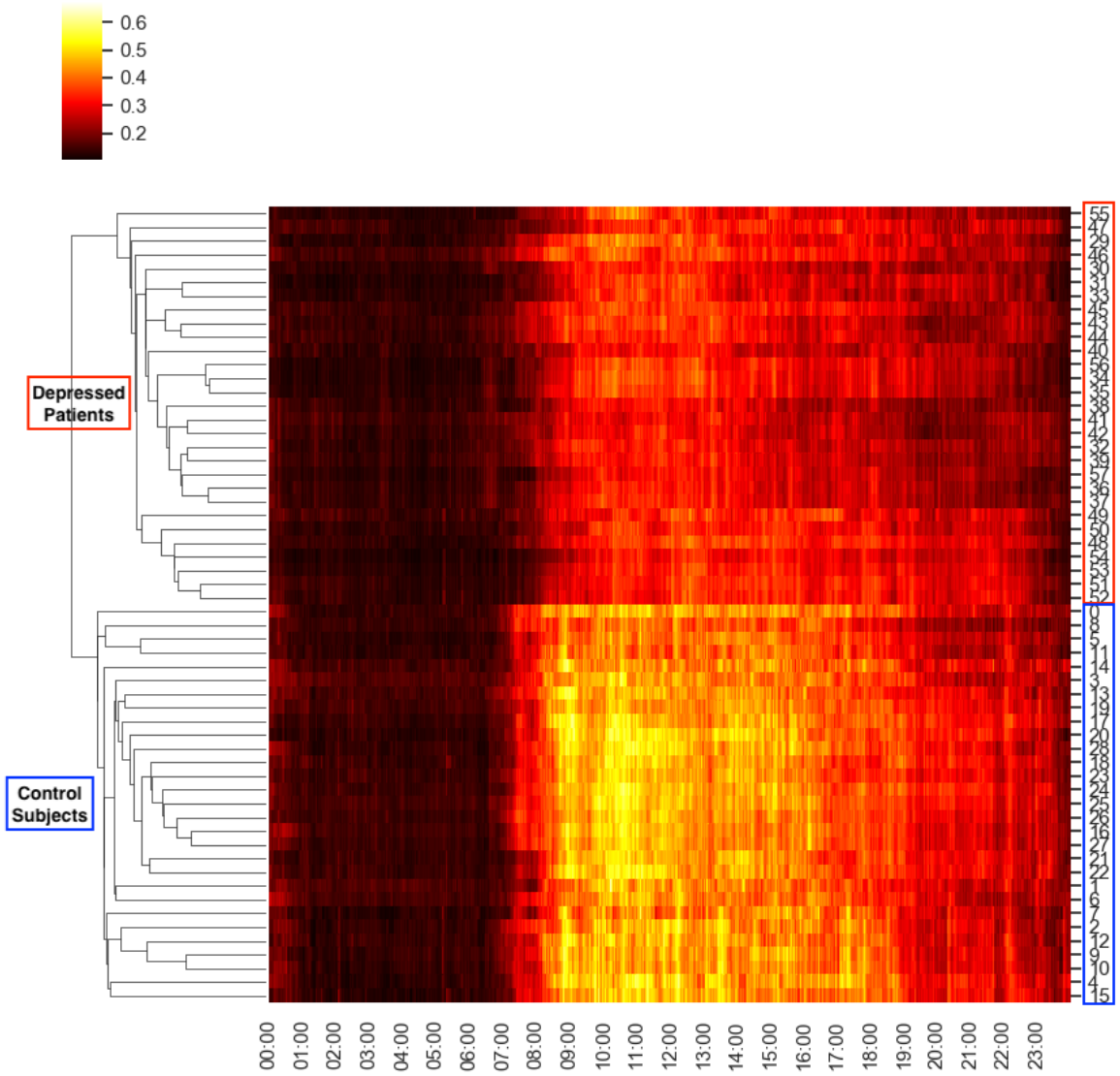

Fig. 3.10 Heat maps for the average of predictions speech in 24-hours.

Table 3.8 Demographic information, clinical and social characteristics, speech detected, and group comparisons.

	LLD patients			**Healthy Controls**			**Statistic**	**Analysis**	
	Mean	**SD**	**n**	**Mean**	**SD**	**n**		**DF**	**p**
Demographics									
Female	72%		21/29	76%		22/29	$\chi^2 = 0.090$	1	0.764
Lives alone	59%		17/29	38%		11/29	$\chi^2 = 2.486$	1	0.115
Age (in years)	74.2	6.0	29	74.0	5.9	29	t = -0.110	56	0.913
NART IQ Score	109.0	9.9	29	114.2	10.3	29	t = 1.962	56	.055
Education (in years)	10.9	2.2	29	13.3	3.3	29	U = 254.0, z = -2.634		0.008*
MMSE Score	27.7	1.8	29	28.8	1.1	29	U = 270.5, z = -2.548		0.014*
Clinical and Social Characteristics									
MADRS (baseline)	28.3	8.6	28	1.0	1.6	29	U = 812.0, z = 6.589		$<$ .001*
GDS-15 (baseline)	9.7	3.1	28	0.9	1.4	29	U = 803.0, z = 6.434		$<$ 0.001*
SF-36 total	38.3	19.9	29	78.7	10.5	29	t = 9.661	42.44	$<$ .001*
IADL total	6.9	1.7	29	8.0	0.2	29	U = 227.5, z = -3.905		$<$ .001*
LSNS-R total	25.5	7.6	29	36.8	7.4	29	U = 123.5, z = -4.625		$<$ .001*
UCLA-LS total	18.5	7.3	29	5.8	4.6	29	U = 765.0, z = 5.366		$<$ .001*
Proportion of Speech detected									
in 24-hours	0.02	0.01	29	0.13	0.03	29	U = 0.000, z = -6.541		$<$ .001*
in the morning	0.04	0.02	29	.24	.04	29	t = 19.980	39.82	$<$ .001*
in the afternoon	0.03	0.02	29	.24	.05	29	U = 0.000, z = -6.540		$<$ .001*
in the evening	.01	.01	29	0.04	0.02	29	t = 8.585	31.05	$<$ .001*

* significant at 0.05 level. LLD=Late-Life Depression; SD=Standard Deviation; DF=Degrees of Freedom

Figure 3.11 illustrates the average percentage of speech detected in a day averaged over the whole week. As we can see from the figure, groups differed in the average quantity of speech detected over a 24-hour period: on average, speech was detected for 2% of the day in LLD, whereas in controls, speech was detected for 13% of the day. This difference was highly significant, and strikingly there was no overlap between groups on this measure. Figure 3.12 shows the mean speech levels for LLD and control groups over a 24-hour period. Groups differed in the quantity of speech detected at each time of day (morning, afternoon, and evening; see Table 3.8). The correlations between mean percentage of total speech detected in 24 hours averaged over the whole week for healthy controls and LLD patients and MADRS, APS, DSSI and LSNS-R are shown in Fig. 3.13, 3.14, 3.15 and 3.16. As illustrated in the figure that the speech is correlated with APS. For LLD, amount of speech detected in total in a 24-hour period was significantly correlated with Attention and Psychomotor Speed, r=0.428, p-value=0.021, where more speech detected was associated with better Attention and Psychomotor Speed.

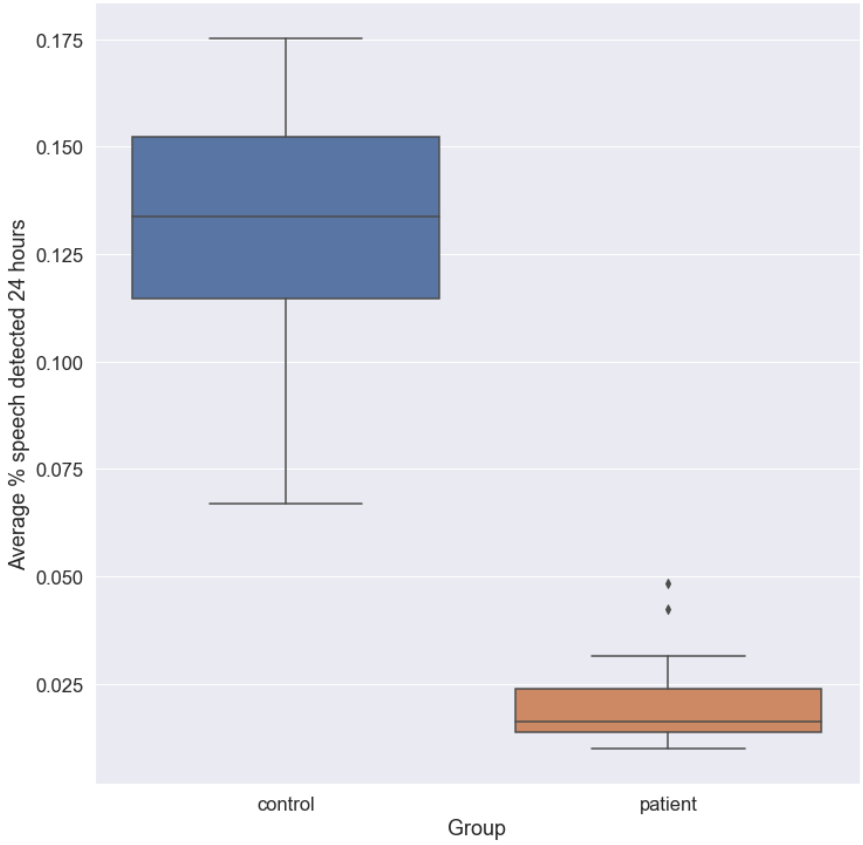

Fig. 3.11 Mean percentage of speech detected in a 24 hour period (averaged over 7 days), statistic=21.504, $p - value < 0.001$

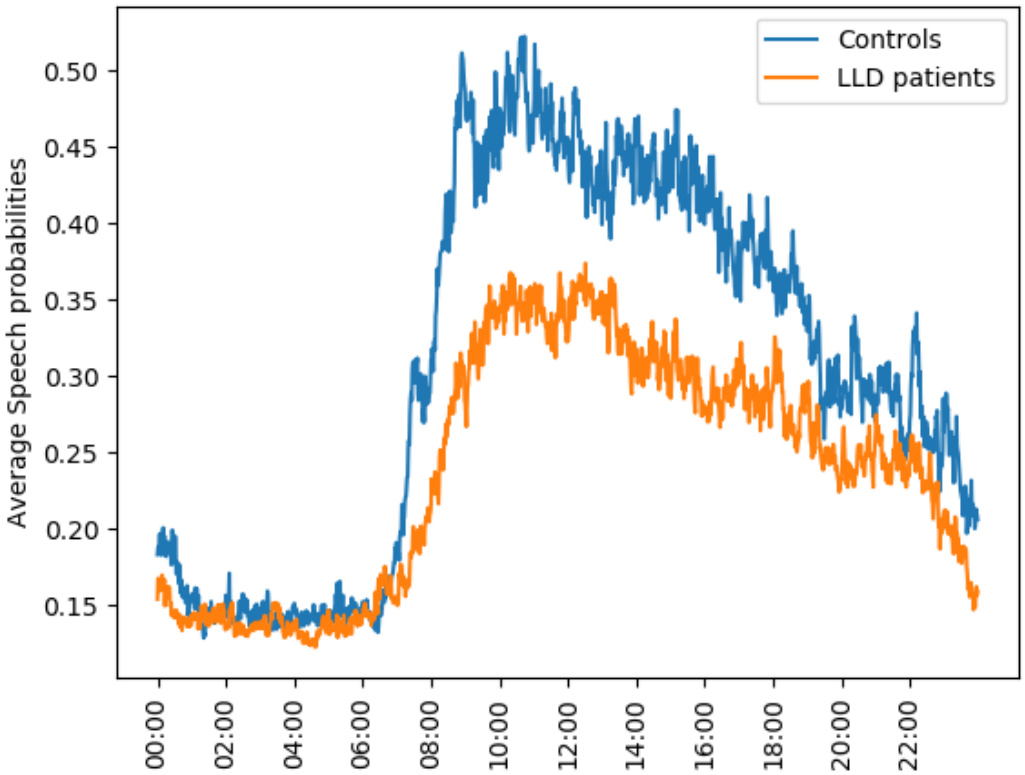

Fig. 3.12 Mean probability of speech being detected for participants with Late-Life Depression (LLD) and healthy controls across a 24-hour period (averaged over 7 days).

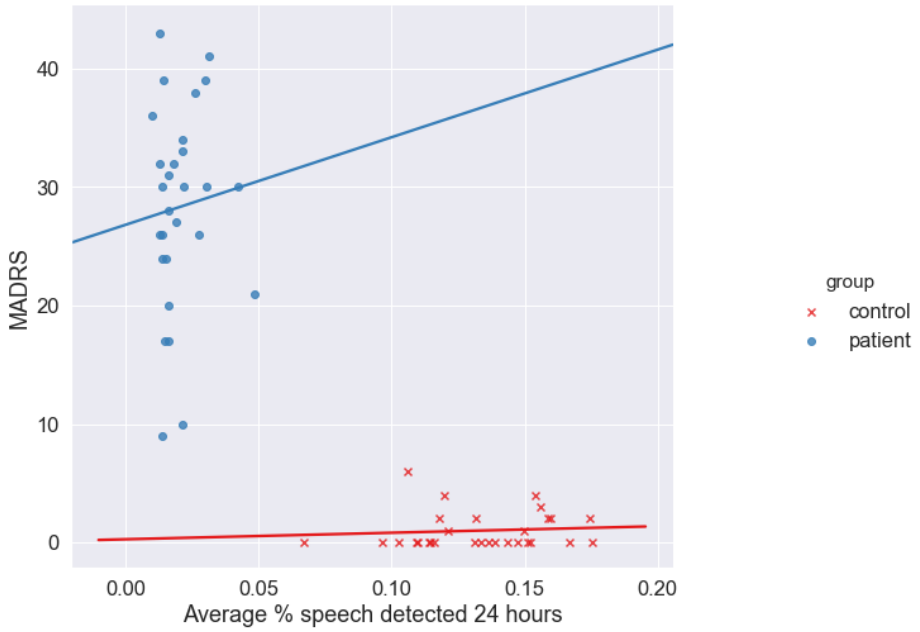

Fig. 3.13 Correlations between average percentage of speech in 24 hours (averaged over the week) with MADRS

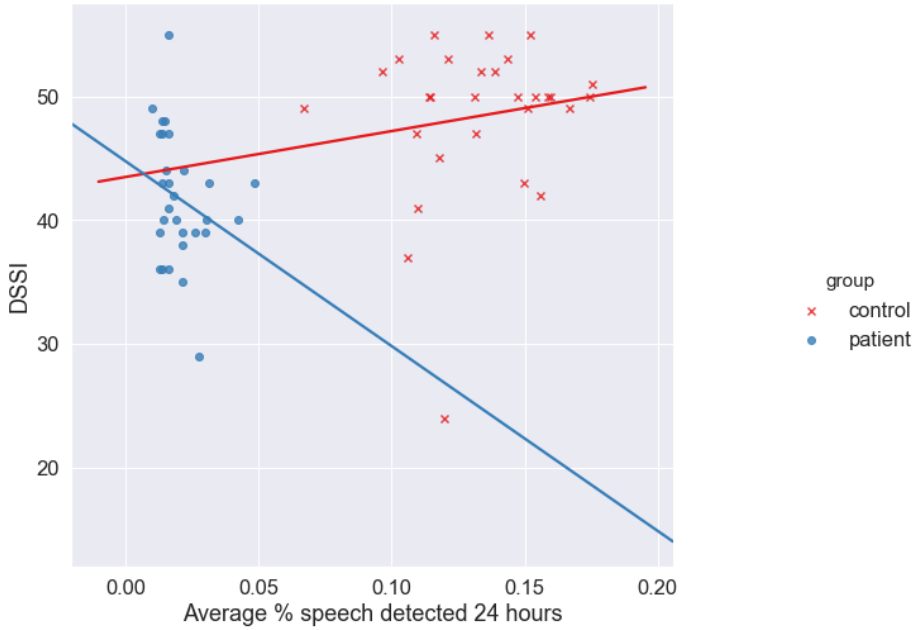

Fig. 3.14 Correlations between average percentage of speech in 24 hours (averaged over the week) with DSSI

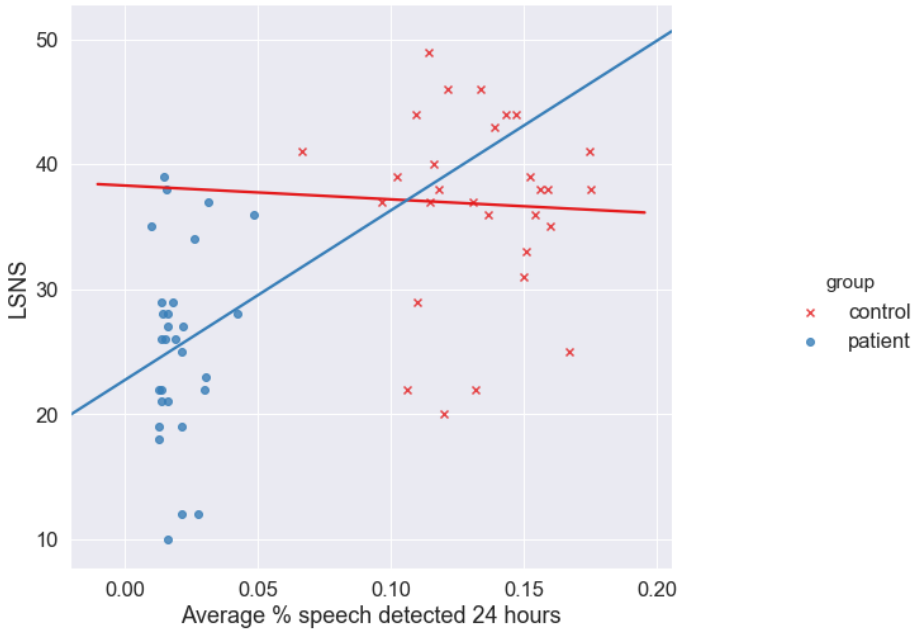

Fig. 3.15 Correlations between average percentage of speech in 24 hours (averaged over the week) with LSNS_R

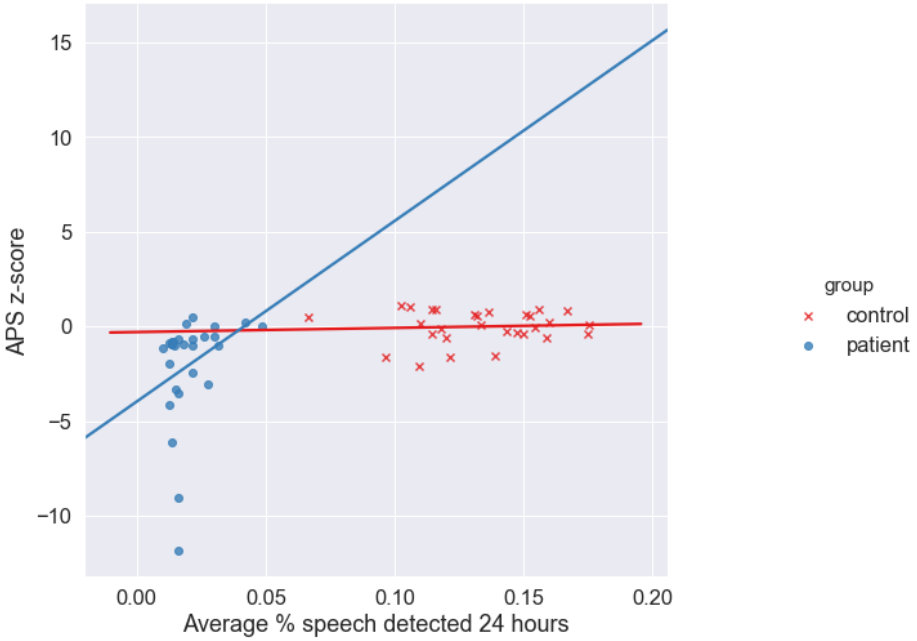

Fig. 3.16 Correlations between average percentage of speech in 24 hours (averaged over the week) with APS z-score

3.5 Conclusion

In this chapter, a deep learning model based on the bi-directional LSTM with gated activation unit, that has been designed to better perform speech prediction from audio data in naturalistic environments. The proposed method is called NatSpeech. The novelty in this approach is feeding the speech to the gated activation unit and then merged together with element-wise sum process. The model is applied to the challenging privacy sensitive data collected in naturalistic settings in order to use the speech detected as a measure to separate between the control and depressed-cohort groups and shows a good correlation with Attention and Psychomotor Speed. Likewise, the model is evaluated using the public datasets (Aurora 2, Aurora 4, TIMIT) and compared to recent literature. From the results and different evaluations, the proposed approach has the capability of separating the two groups of healthy controls and patients with depression in addition to its superior performance in the case of the public datasets.

Chapter 4

Wearer Speech Detection as a Biomarker for Depression

4.1 Introduction

4.1.1 Background & Motivation

Speaker speech analysis is a demanding field in various applications such as speaker segmentation or diarization [89], speaker verification [20], and speaker recognition [17]. Speaker speech analysis refers as to the learning of the speaker speech characteristics. This analysis can be used effectively for finding the speaker turn taking and then identifying the speaker in the conversation and this answer the second research question of "who is talking to whom and when". That means given an audio signal or even video recording, the task is to determine speaker turn taking within a conversation of unknown number of people. Speaker speech analysis is relying definitely on the detection of speech and non-speech segments from the signal. The conventional approach shown in Fig. 4.1 for speaker speech analysis is starting with preprocessing the audio signal and extracting the speech features to be used for segmenting the audio signal into speech and non-speech segments. The idea beyond the feature extraction is a comparatively low dimensional and less redundant representation or manifold to be used further for statistical modelling. However, these features are focusing on some characteristics of the audio signals and then loses some information and therefore fail to capture the speaker characteristics especially in noisy or naturalistic-setting environments. Examples of features includes but not limited to Mel-Frequency Cepstral Coefficients (MFCC) [51], Perceptual Linear Predictive Coding (PLP) [65] and Linear Predictive Coding (LPC) [9] as discussed in Chapter 2. Following the feature extraction is the speech detection model and then the change point detection which relies on the unsupervised learning approach (clustering) to

cluster the speaker in the conversation. The state-of-art unsupervised learning are often based on measuring the statistical distance between two consecutive frames in the speech signal.

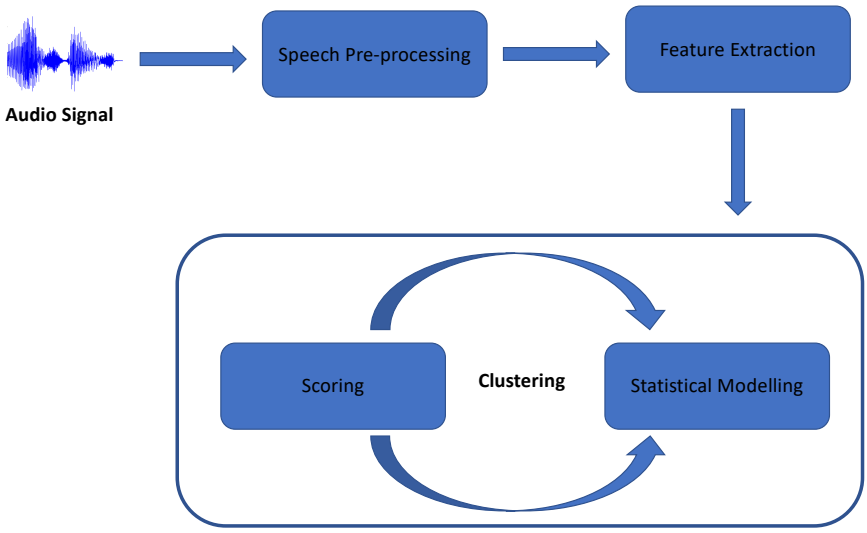

Fig. 4.1 Conventional Speaker speech analysis.

In order to solve the issue of the feature extraction, many researchers have have relied on deep learning models. DNN or fully connected networks and auto-encoders [54] showed a superior performance for feature extraction: these features represents the main characteristics of the speaker speech. The encoding layer or the bottleneck layer of the autoencoders has been utilised as the latent representation vectors or the speaker embeddings and has been used for speaker segmentation, diarization or classification applications [70],[137]. Prince et al. [137] generated the speaker embeddings using a supervised classification DNN model and they utilise the layer before the softmax layer of DNN for that purpose by optimising the speaker misclassification error. These embeddings generated from DNN are exploited to detect the speaker turn taking points and then speaker segmentation and clustering. Another example of utilising deep learning models for learning speaker speech characteristics is in [21] where Chen and Salman used a loss function that is based on whether the input speech frame belongs to the same speaker or not. Also in [184], multiple autoencoders architectures is leveraged for extracting the speaker speech for the purpose of speaker diarization.

4.1.2 Approach

In this chapter, the aim is to infer how much and how often the wearer of the WAM device (discussed in Chapter 3) engage in any conversation and how many conversations and the conversation duration. What is different here from the state-of-art methods that the task is mainly to find the main speaker within the detected speech, specifically detects the person who wears the WAM device and consider any other speech detected is considered as "other speaker". This will answer the question of how much the wearer engage in any conversation. So, basically, the goal is to detect the wearer from the speech that is already detected by NatSpeech method discussed in Chapter 3. In order to achieve this goal, the raw audio frames will be fed to the speech detection model to detect the speech segments and then these speech segments will be passed to another model to detect whether the speaking person is the main speaker (wearer) or not. Therefore, I will present another deep learning-based wearer detection method for the same challenging circumstances of naturalistic acoustic environments. For the sake of this, an end-to-end speaker speech detection model in naturalistic environment conversation for elderly depressed people is introduced.

The model is based on convolutional autoencoder used as pre-training to bidirectional LSTM followed by LSTM layers mixed with maxout layers and Leaky ReLU layers. The architecture is quite similar to NatSpeech discussed in Chapter 3. The difference is that the architecture is using the convolutional autoencoder rather the gated layers. The input to the network are the normalised raw frames of speech data without any explicit feature extraction/engineering and followed by evaluation on the collected datasets (Discussion dataset and Depression dataset) that have been generated by WAM device and used in the previous chapter. The Discussion dataset is used for training and validation purposes and the Depression dataset is used for the testing purpose. I first evaluate the capacity of the method to successfully predict the wearer on the discussion dataset. Thereafter, based on the transfer learning concept discussed in Chapter 3, the proposed wearer detection method can be effectively trained on the discussion dataset and then used to annotated the depression dataset. Furthermore, this method can be used as a digital marker for depression and some statistical analysis that includes the significance of the digital marker and the correlation with key clinical variables is also presented in Section 4.3.

4.2 Method

In this section, a detailed overview of convolutional autoencoders will be presented along with the network structure that shares part of the previous structure discussed in Chapter 3.

4.2.1 Convolutional Autoencoders

Convolutional autoencoders (CAEs) [40] comprise two main subnets, the encoder subnet and the decoder subnet. The encoder subnet takes the audio data in my case after some preprocessing and maps these preprocessed data into bottleneck features or the bottleneck latent representation. The decoder subnet will then try to reconstruct the input back from the the bottleneck representation. The encoder subnet in CAEs consists of a convolutional layer, a pooling layer and often fully-connected layer for the bottleneck features as shown in Fig. 4.2. The convolutional layer is composed of a kernel that convolve over the input data to make the receptive field as discussed in Chapter 2 and a set of filters that are able to form the feature map. The pooling (aka. the subsampling layer) to downsample the latent representation based on the kernel size and apply (maximum or average function) to achieve this. The fully-connected layer is optional since the main reason for using it is its high efficacy for the purpose of classification and in this thesis only one-fully-connected layer is used. The aim of the encoder is to form a better representation from the input and then passes it to the decoder. The decoder objective is then to reconstruct the original input and based on the loss function the parameters are updated until there is no change in the reconstruction error. The most conventional loss functions are reconstruction cross entropy, Kullback-Leibler (KL) divergence and mean square error (MSE) [54] and the one used in this thesis is the reconstruction cross entropy. The decoder subnet comprises repetitions of a convolutional, activation layer, an up-sampling and fully-connected layers. The upsampling in this case is trying to do the opposite of the pooling layer by repeating the activations according the upsampling kernel size and then the convolution process will update the filters until it reaches the reconstruction layer. The process of upsampling with the convolution is called the deconvolution process.

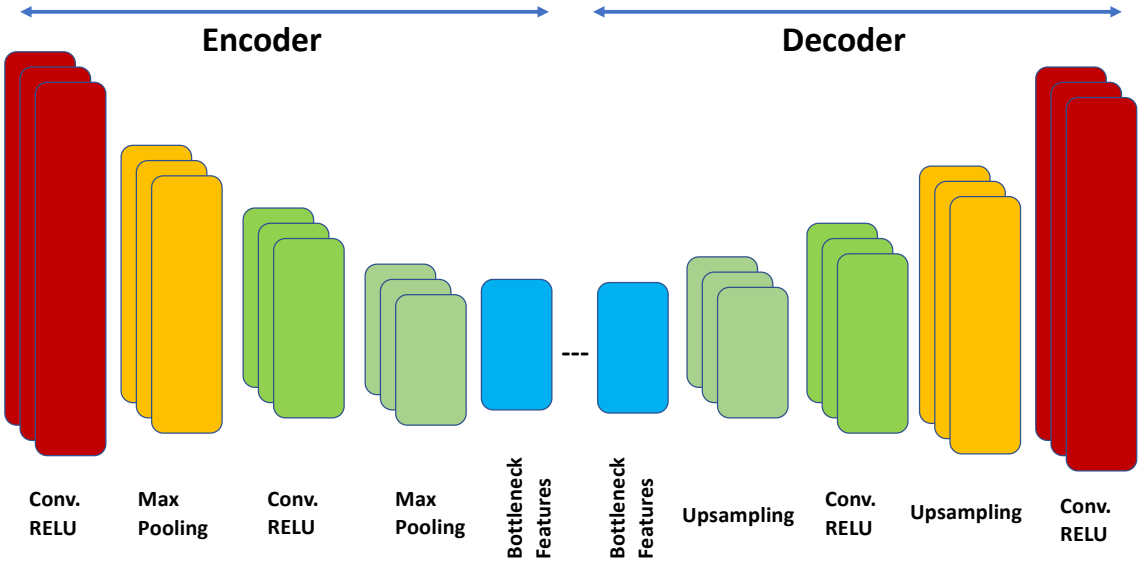

Fig. 4.2 Convolutional Autoencoder Architecture.

4.2.2 NatWearer: CAE-BLSTM

The proposed model (called NatWearer) for detecting the wearer is composed of CAE and Bi-LSTM layers. The basic idea of CAE part is to extract the latent features of the speech produced by the wearer and non-wearer and Bi-LSTM will then take care of the past history of the wearer's speech and also the future context to help make better predictions. The preprocessed audio frames after normalising it as in the Chapter 3 are passed to the speech detection model to identify the speech frames first and only the speech frames are passed to NatWearer presented in Fig. 4.3. The predicted speech frames which are already normalised are then fed to the CAE which play the role of pretraining to extract the relevant features of the speakers. The reason beyond CAEs here is that CNNs is proven to do a very powerful feature engineering process within the model. However, the convolutional filters which are responsible for extracting the features and construct the feature map requires to be optimised. CAEs have the capability to allow for the autoencoder model to learn the optimal convolutional filters that fundamentally minimise the reconstruction error [106]. This leads to a better compact representation from the input. This helps to efficiently learn the speaker characteristics that leads the wearer detection model to better performance.

In this chapter, a wearer detection model is presented. This model is different from speaker segmentation or diarization as in the literature. The reason for considering only the main wearer of the WAM device and considering any other speech as non-speaker is that

the aim of the work is to analyse the speech produced by the depressed patient (the wearer in this case) and not to all of the speakers. The hypothesis is that the wearer is closer to the device from the other speakers most of the time. In the training process, the decoder is used to reconstruct the speech frames and then in the validation and test sessions only the encoder is used. Following the CAE, there are four BLSTM layers that leverages the dropout and gradient normalisation as a regularizers followed by six layers of LSTM with dropout and l_2 regularizer as a penalty to the loss function. These set of RNN layers that includes BLSTM and LSTM layers are basically used to learn the temporal structure of the main speaker (who wears the device) and other speakers talking to them. Following the RNN layers is the three maxout layers and one fully connected layer with Leaky ReLU as an activation which then passes the output to the softamx for the final classification of speech frames to main wearer and other speaker.

The difference between NatWearer and NatSpeech (the model used for the speech detection) is that the gated layers is replaced by CAE and also the LSTM layers are used with a penalty on the loss function (l_2 regulariser). The reason for that is the decline in the performance of NatSpeech to identify the main wearer in the training dataset. As discussed in Chapter 3 where the justification of each part of the network was presented. The difference here is that the model needs to make a latent space for each speaker after the speech is already detected. In this case the gated layers will not help and autoencoder will be the best option. The choice of CNN layers for autoencoder is that CNN does not require much time in the training compared to LSTM. The comparison of the evaluation between NatSpeech and NatWearer model to detect the main speaker is justified in Section 4.3. The source code of CAE-BLSTM method is available in the Java class *NatWearer.java* at my GitHub https://github.com/OssamaAlshabrawy/NatSpeech.

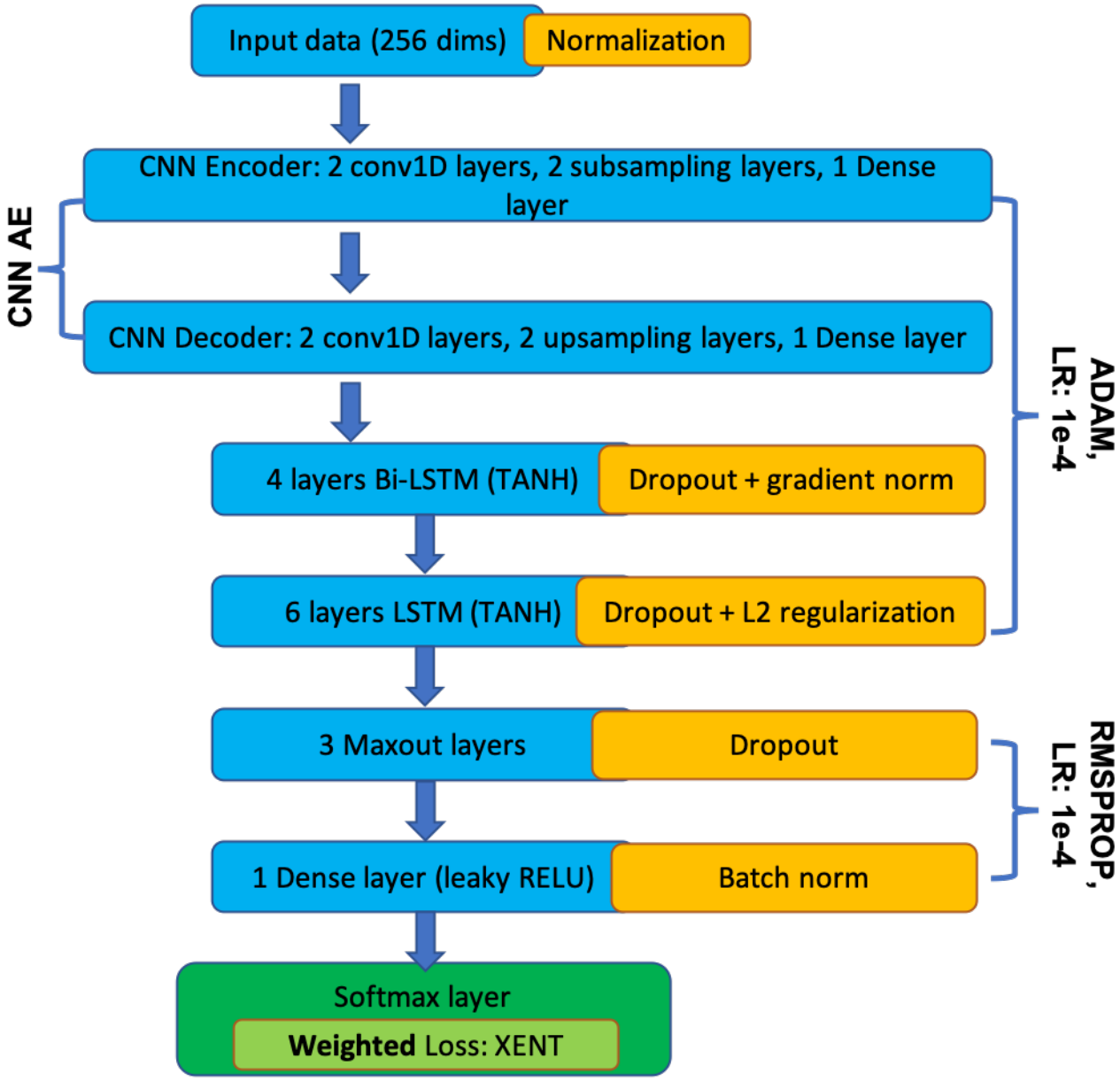

Fig. 4.3 Wearer Detection Model Architecture.

4.2.3 Network Training

The training phase is performed by backpropagation and due to the different structures of the layers, different types of propagation is used. For the CAE, the weights are updated using the backpropagation like in MLP but the filters are updated using the convolution process. For BLSTM and LSTM layers backpropagation through time (BPTT) is used and the optimisation algorithm used is Adam. The loss function utilised for the whole model is the cross-entropy. The learning rate used for Adam is 10^{-4} unlike what is used in NatSpeech which was 10^{-5}.

The learning rate is chosen based on grid search with the values $(10^{-1}, 10^{-2}, ..., 10^{-6})$ as discussed in Chapter 3. The final four layers (maxout and dense) are trained using RMSPROP and the learning rate is 10^{-4} exactly like in NatSpeech. Over-fitting of the training set is managed by early stopping algorithm where the maximum number of epochs is set to be 100. For parameter regularisation, L_2 regularisation is used. Nevertheless, the dropout exploited in NatWearer has only 0.5 retaining probability. The dropout is chosen using a grid search with the retaining values (0.2, 0.3, 0.4, 0.5, 0.6, 0.7 and 0.8). In addition, batch normalisation and gradient normalisation is the same as NatSpeech. The number of LSTM cells in each layer is number between 64 and 256 and optimised by HyperParameterOptimization tool in Deeplearning4j [43].

4.3 Results and Discussion

As stated in the introduction, the state-of-the-art methods are trained on a clean speaker speech with a variety of speakers with annotation to each of these speakers but the training set utilised here only has annotations for the main wearer and all other speakers have annotation of "other speaker". Also their datasets are recorded in lab-settings but the data recorded with WAM device is noisy even the speech segments are not fully clean. The evaluation of NatWearer is performed on the private datasets described in the previous chapter (namely the discussion and depression datasets). As discussed before the depression dataset is unannotated dataset due to the privacy sensitive content of the speech which come out from the recording in unconstrained environment. The unconstrained environment means that the data is recorded freely by using the WAM device according to the preference of the participant and use it in the everyday life activity without any constraints. This points out that the device is unobtrusive to the participants. So, the challenge in the Depression dataset is not only the data is unannotated but also using the device everywhere without constraints will create sophisticated indoor scenarios such as audible speech within background which includes people speaking in the background, TVs, or any device that generate human speech. Likewise, those constraints creates sophisticated outdoor scenarios like different types of noises from passer by, streets and vehicles.

In order to evaluate the depression dataset in terms of using the amount of wearer's speech as a digital biomarker for depression and find the association with the clinical measures related to depression, I evaluate the performance on the discussion dataset and also perform all the training and validation on it since the discussion dataset is labelled. This applies the concept of transfer learning as discussed in Chapter 3. Table 3.2 shows some statistics for the sessions of the Discussion dataset with details about each session. The first column represents

the session number, the second represents the circumstances of the recording environment, and the third and fourth column represent the percentage of speech produced by the wearer and other speakers respectively. This shows the circumstances of each recording whether it is recorded indoor or outdoor and the place of recording if possible and the presence of noise or the audible speech within background. In the following evaluations, I will show the classification performance using NatSpeech first and the motivation to further development of the model and end up with NatWearer system.

Table 4.1 Performance evaluation measures of the CAE-BLSTM method on the discussion dataset using NatSpeech system

Session	Accuracy	Precision	Recall	F1-Score
1	0.9	0.91	0.9	0.9
2	0.81	0.82	0.81	0.81
3	0.87	0.90	0.87	0.87
4	0.80	0.87	0.80	0.80
5	0.88	0.91	0.88	0.88
6	0.86	0.88	0.86	0.85
7	0.89	0.91	0.89	0.89
8	0.75	0.88	0.75	0.77
9	0.90	0.96	0.90	0.92
10	0.96	0.95	0.96	0.96
11	0.91	0.90	0.91	0.91
12	0.87	0.89	0.87	0.86
13	0.87	0.91	0.87	0.89
14	0.90	0.93	0.92	0.93
15	0.95	0.98	0.97	0.96

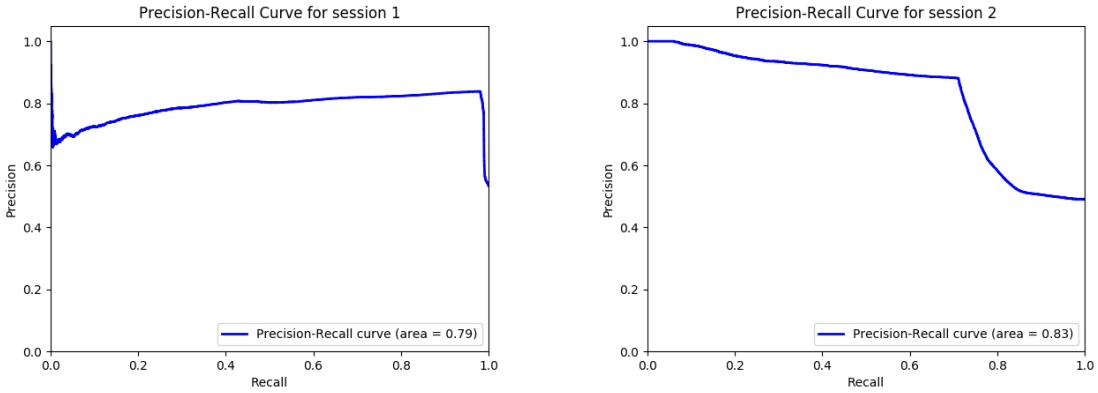

Fig. 4.4 PR curves for session 1 and session 2(Discussion Dataset)

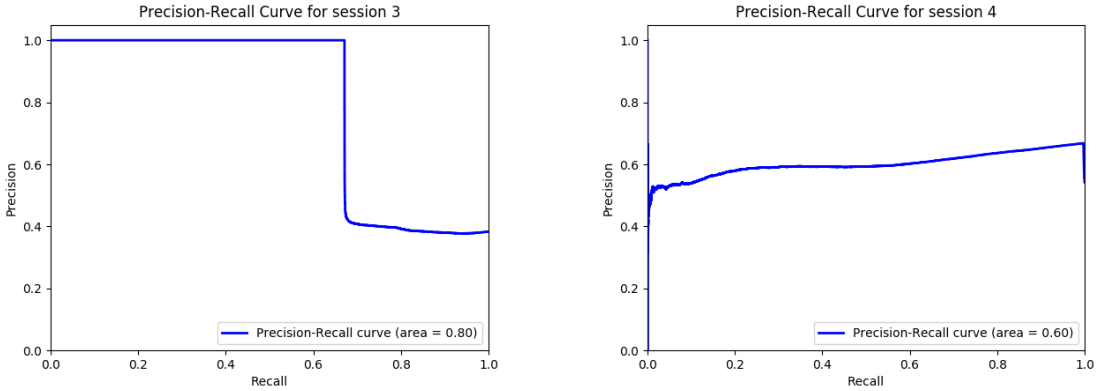

Fig. 4.5 PR curves for session 3 and sessions 4 (Discussion Dataset)

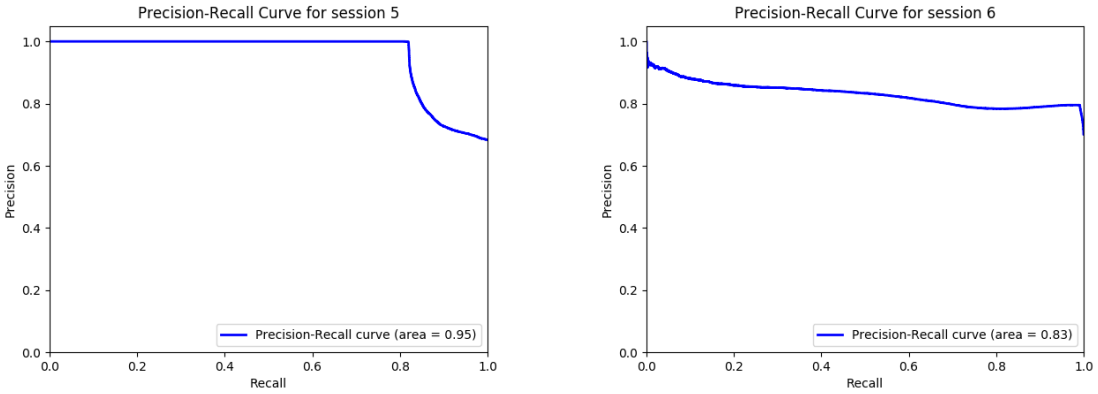

Fig. 4.6 PR curves for session 5 and session 6 (Discussion Dataset)

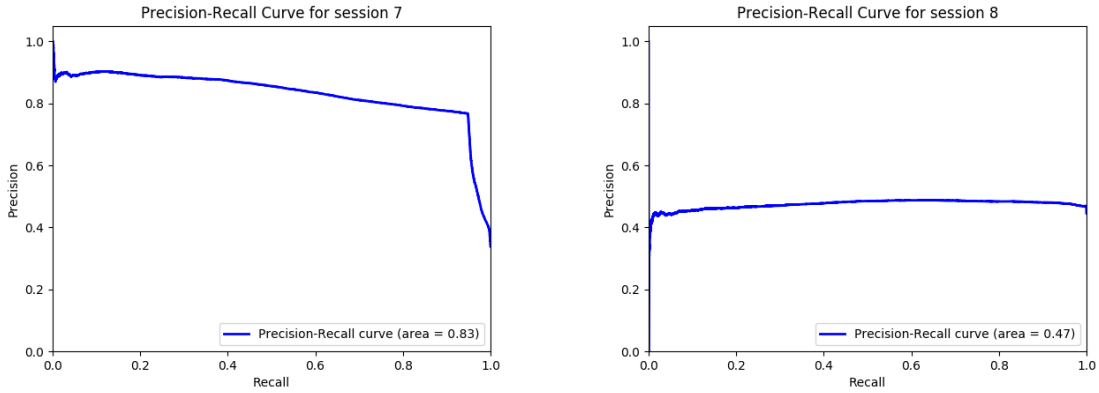

Fig. 4.7 PR curves for session 7 and session 8 (Discussion Dataset)

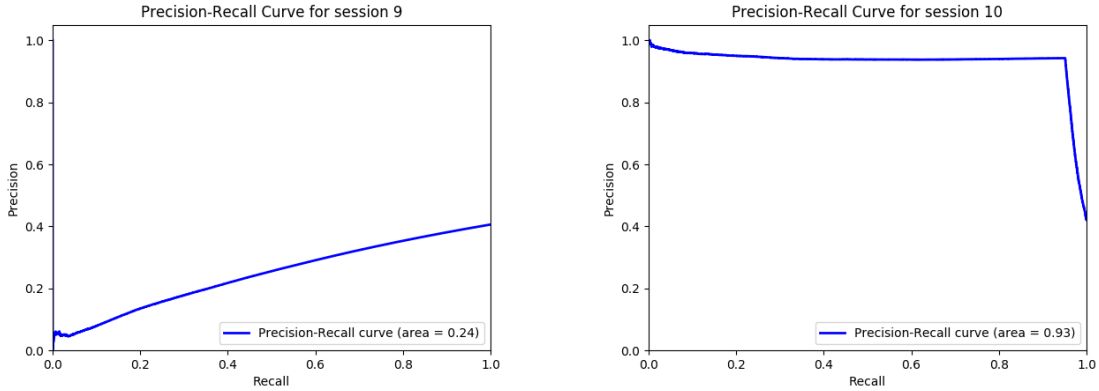

Fig. 4.8 PR curves for session 9 and session 10 (Discussion Dataset)

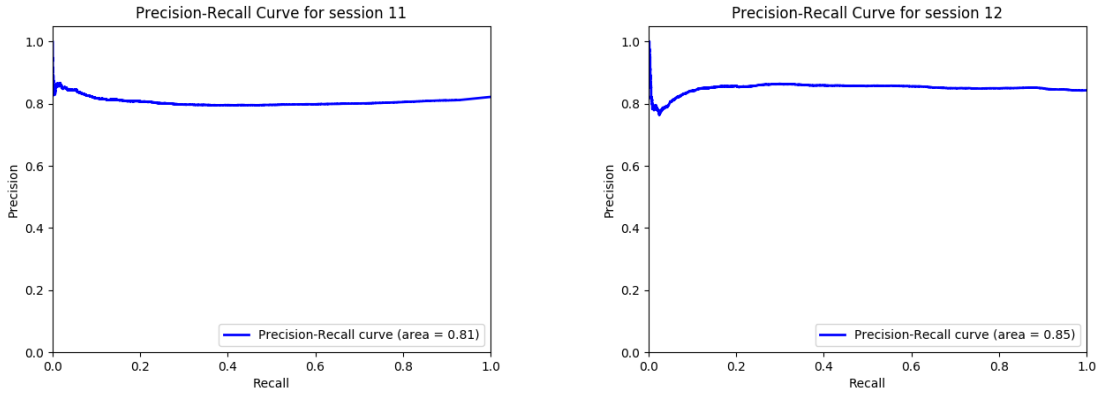

Fig. 4.9 PR curves for session 11 and session 12 (Discussion Dataset)

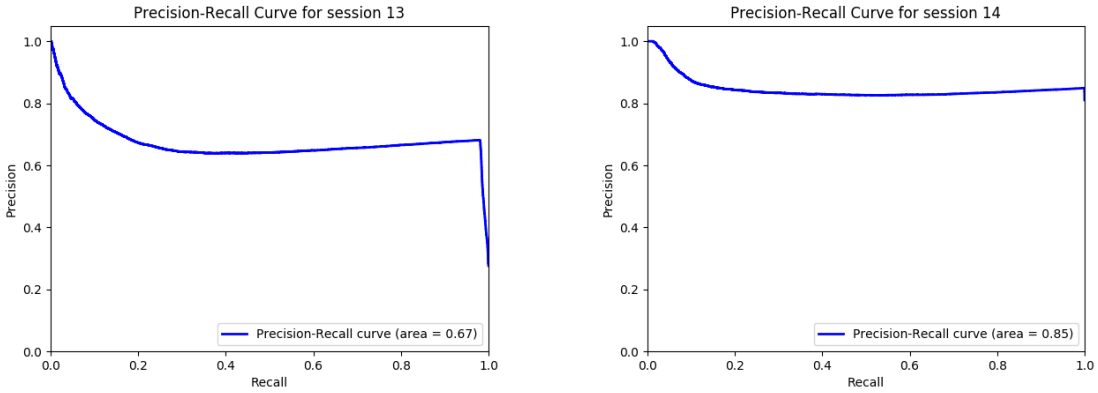

Fig. 4.10 PR curves for session 13 and session 14 (Discussion Dataset)

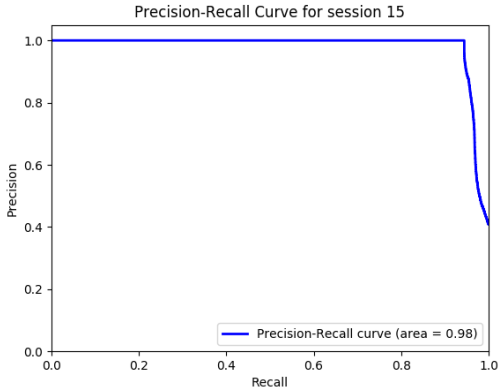

Fig. 4.11 PR curve for session 15 (Discussion Dataset)

Table 4.1 shows the evaluation on the discussion dataset using NatSpeech model that has been used in the previous chapter. The results show a decline in the performance in terms of accuracy, precision, recall and F1-score compared to the performance obtained for speech detection previously. The decline is not that high, however, we care more about the positive class which in this case will be detecting the wearers to infer the amount of speech produced by them. In order to justify that, precision-recall curves (PR-curves) are used whereas ROC-curves (made up of recall and specificity) are looking at both classes. The Precision, recall, F1-score are discussed in Chapter 2 Section 2.5.

The PR curves show that the algorithm struggle in the sessions 4, 8, and 13 (Figures 4.5, 4.7 and 4.10) with area under curve of 0.60, 0.47, 0.67. The algorithm completely struggle in session 9 with only 0.24 area under curve. This could be due to the fact that those session are outdoor sessions with too much noise except session number 8 which is indoor but in noisy room which makes the speech mixed with a lot of noise coming out of the room itself

and from the background as well. Also, session 9 (Figure 4.8) has only 10% of wearer's speech. Although the F1 score shows a quite good performance for the most of the sessions, our interest is the main wearer at the end. So, the PR curves will be used in the final decision. This is the justification why there is another model for detecting the wearer. In the next evaluations, I display the results of NatWearer and how it can produce a better predictions for the wearer through the sessions of the Discussion dataset.

Table 4.2 Performance evaluation measures of the CAE-BLSTM method on the discussion dataset using NatWearer system

Session	Accuracy	Precision	Recall	F1-Score
1	0.91	0.9	0.92	0.91
2	0.95	0.93	0.98	0.95
3	0.91	0.94	0.89	0.9
4	0.87	0.89	0.87	0.87
5	0.93	0.92	0.93	0.92
6	0.9	0.9	0.91	0.91
7	0.92	0.9	0.93	0.91
8	0.91	0.86	0.94	0.89
9	0.95	0.97	0.95	0.96
10	0.97	0.97	0.96	0.96
11	0.96	0.95	0.95	0.95
12	0.91	0.89	0.91	0.9
13	0.92	0.89	0.94	0.91
14	0.93	0.92	0.93	0.92
15	0.98	0.98	0.97	0.97

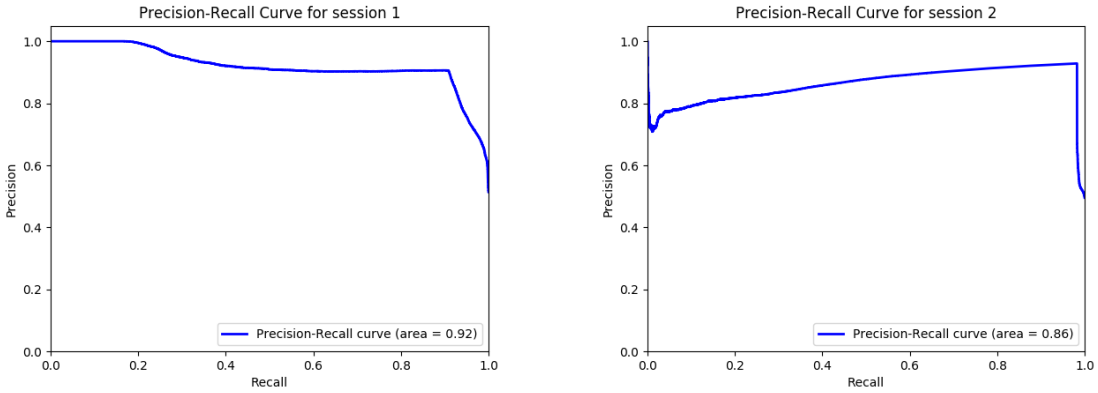

Fig. 4.12 PR curves for session 1 and session 2(Discussion Dataset)

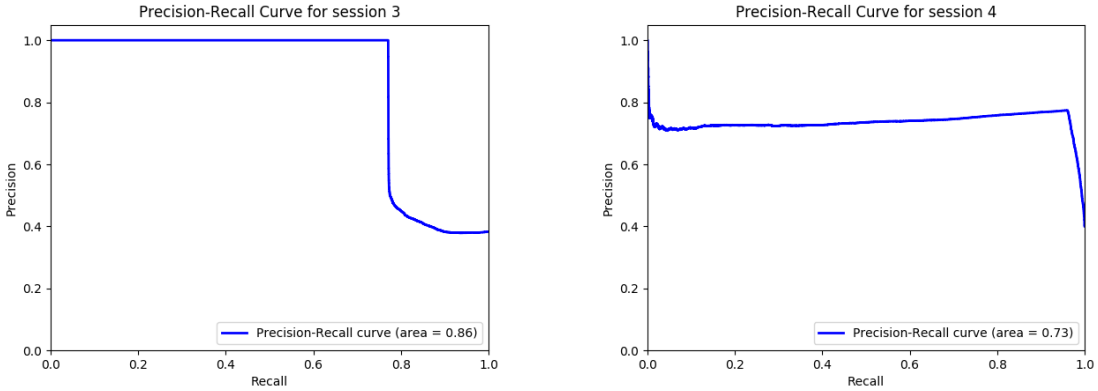

Fig. 4.13 PR curves for session 3 and sessions 4 (Discussion Dataset)

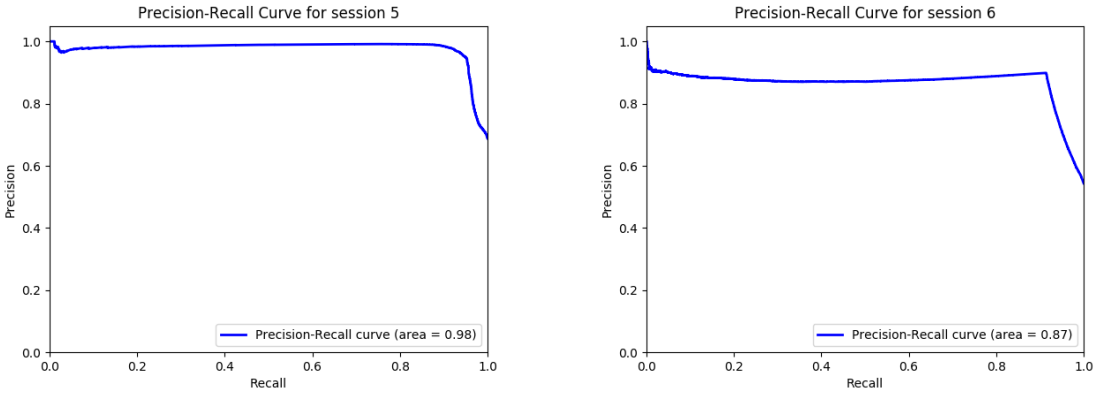

Fig. 4.14 PR curves for session 5 and session 6 (Discussion Dataset)

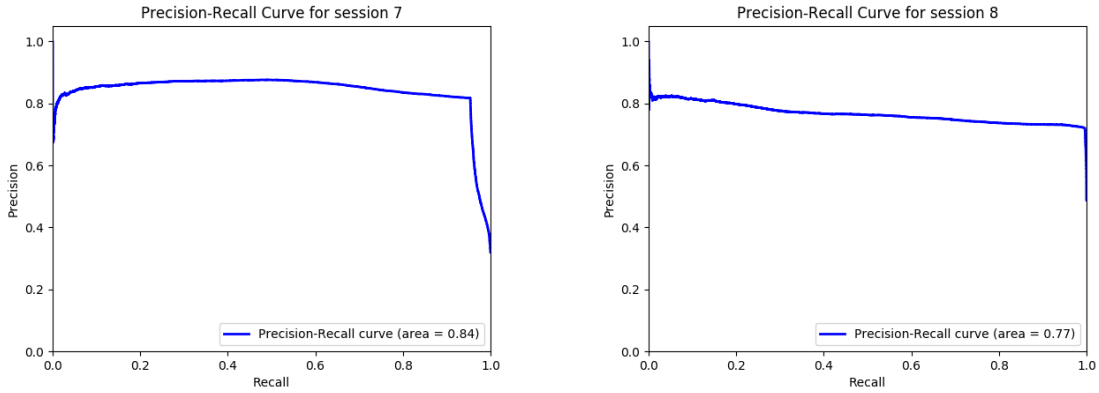

Fig. 4.15 PR curves for session 7 and session 8 (Discussion Dataset)

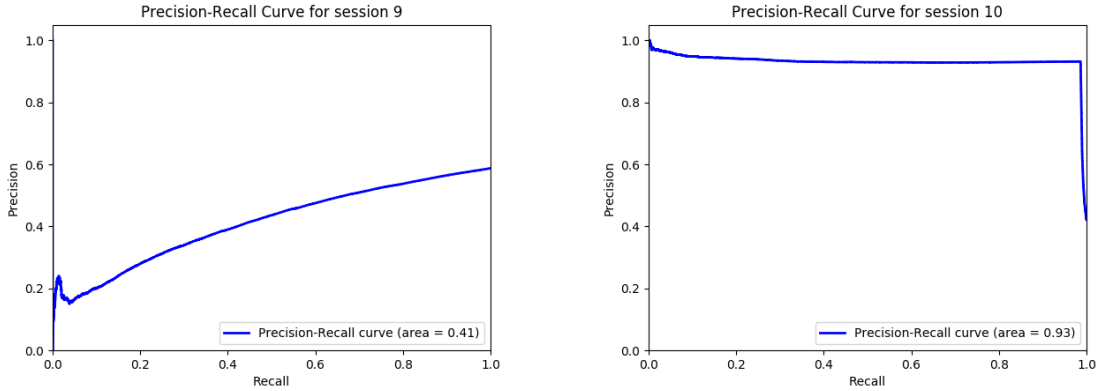

Fig. 4.16 PR curves for session 9 and session 10 (Discussion Dataset)

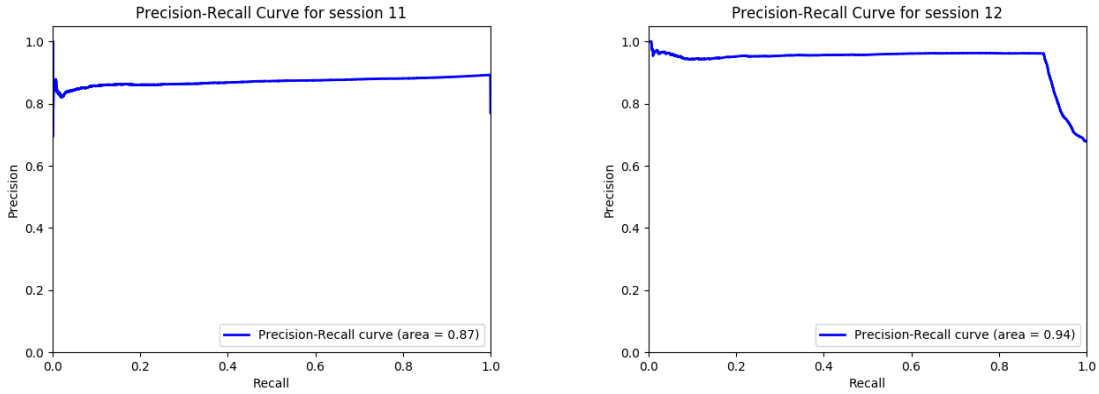

Fig. 4.17 PR curves for session 11 and session 12 (Discussion Dataset)

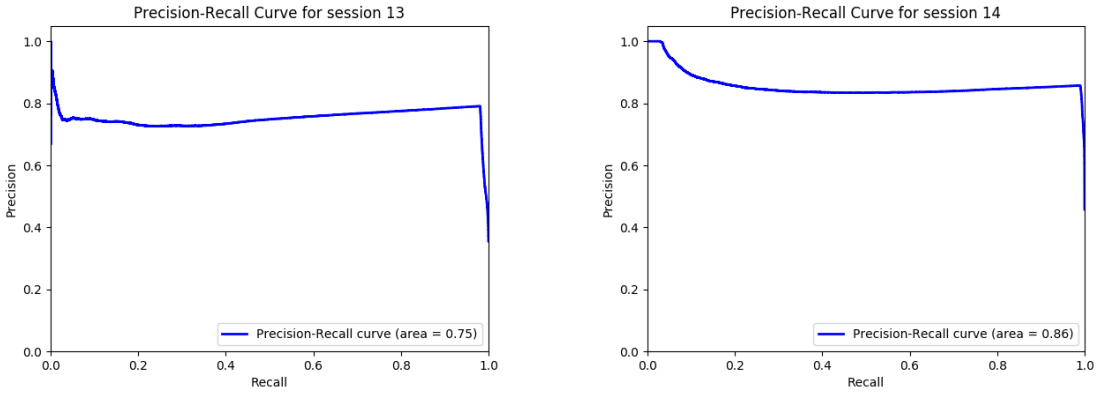

Fig. 4.18 PR curves for session 13 and session 14 (Discussion Dataset)

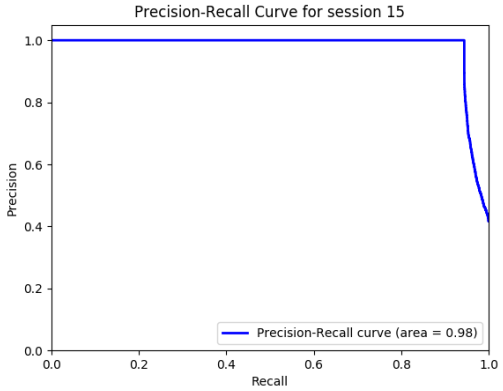

Fig. 4.19 PR curve for session 15 (Discussion Dataset)

Table 4.2 shows remarkable improvements in accuracy, precision, recall, F1-score. Notwithstanding, the area under PR-curve for the session 9 is still low but better than NatSpeech. NatWearer is designed to tackle the challenge of different outdoor noises that could be mixed up with the speech as well as when speech is mixed up with the indoor audible speech within the background. The clue behind observing PR-curves is to investigate how well the classifier performs when predicts the wearer speech and hence this leads to enhance annotating the depression dataset in terms of speech produced by the wearer and then a conclusion can be drawn from the analysis of significance between the control and patient groups.

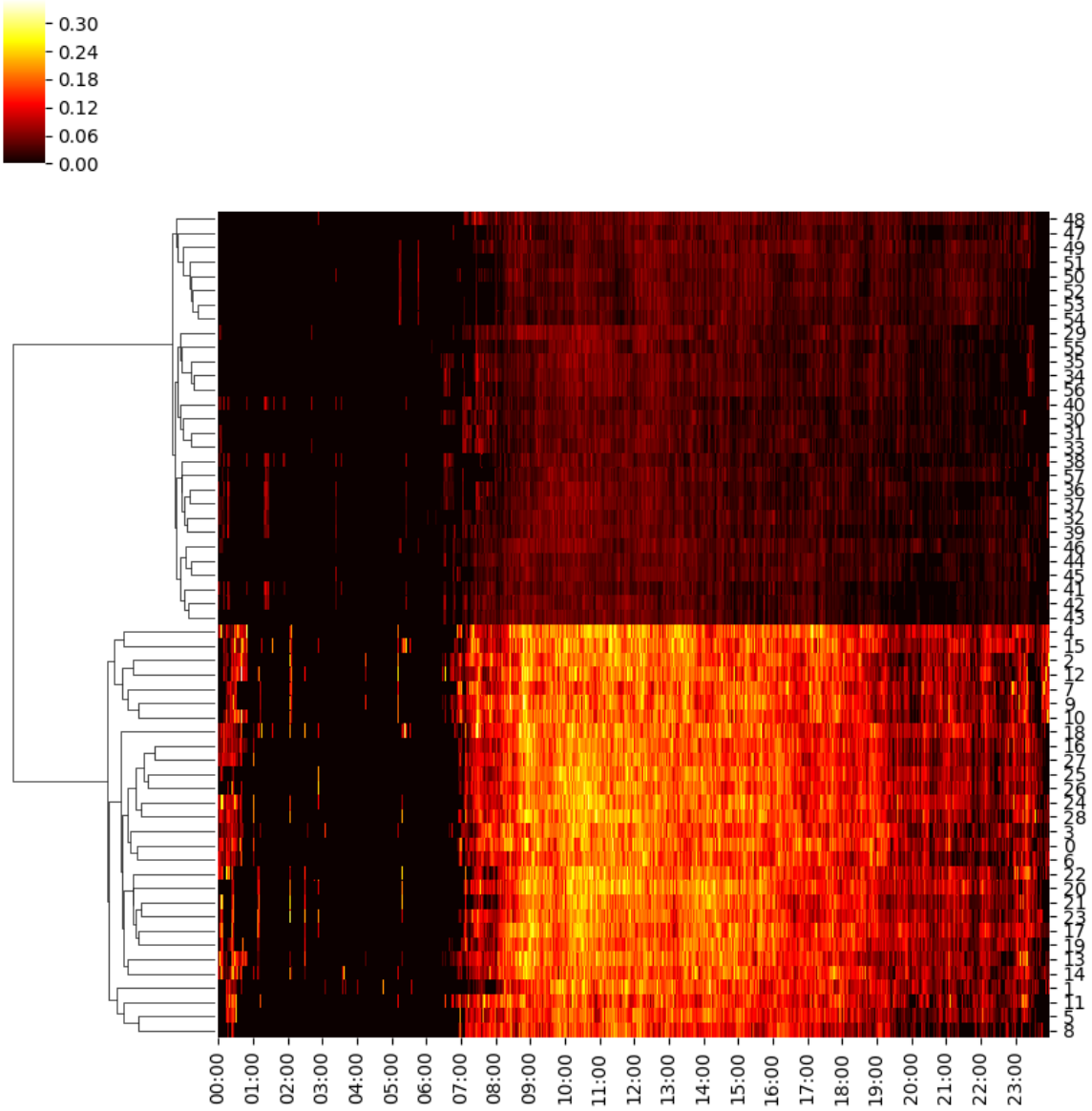

Fig. 4.20 Heatmap of the average prediction of speech produced by the wearer in a 24-hours averaged over the week. On the right side, the numbers from 0 to 28 are the IDs for the control subjects and the number from 29 to 57 represents the LLD patients. On the left side of the graph is the dendrogram that shows two main clusters that perfectly separated. The colour bar in the up-most left represents the percentage of speech produced by the wearer. We could see from the dendrogram that there is no overlap between the groups and they are separated perfectly.

Besides all the evaluations done on the Discussion dataset, the next figures are showing some evaluations and statistical tests on the Depression dataset. The predictions of the wearer's speech in a day averaged over the whole week are displayed in Fig. 4.20. Each

minute of the prediction is the average of all of the corresponding minutes in the week. For instance, the prediction of the minute *9:45* is the average of *9:45* in day 1, *9:45* in day 2, ..., and *9:45* in day 7. The heatmap and the dendrogram show 2 separate clusters (one of the control subjects) and (one for the patients) which highlight the efficacy of wearer's speech as a good digital biomarker for Depression. In order to evaluate the significance of the biomarker, Fig. 4.21 shows the box plots for the two groups. Groups differed significantly in the proportion of speech they produced themselves (out of all speech detected): for LLD, (3 $\pm$ 0.03)% of the speech detected was produced by the wearer; and for controls, (11 $\pm$ 0.01)% of the speech detected was produced by the wearer.

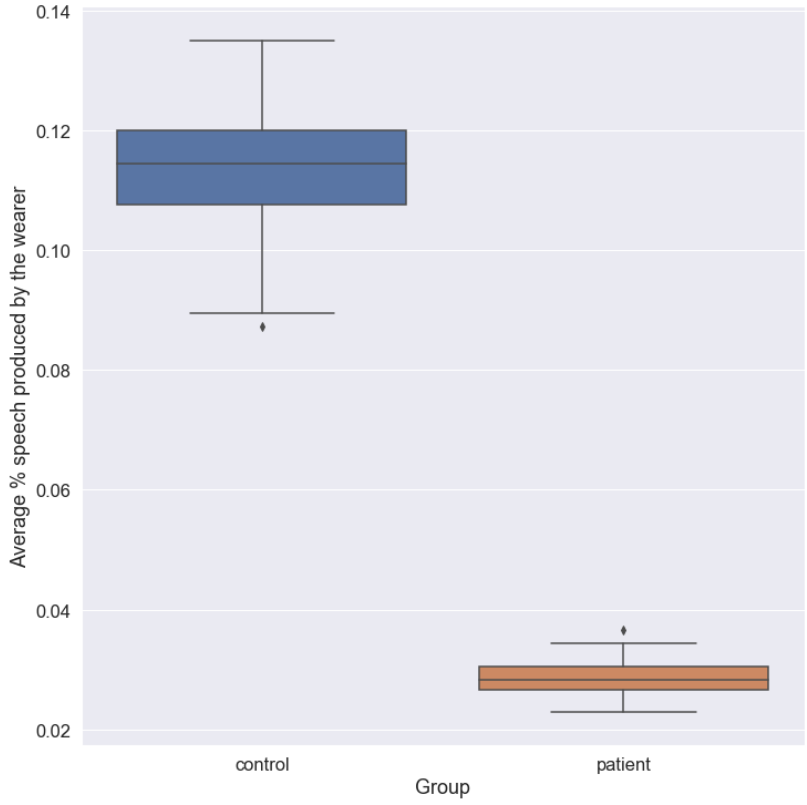

Fig. 4.21 Mean percentage of speech produced by the wearer themselves (out of all speech detected), for Late-Life Depression (LLD; N=29) and healthy controls (N=29). Dots represent individual participants and are randomly spread across the x-axis within each group. Speech detected by the wearers themselves in LLD patients produces a smaller proportion compared to healthy controls ($statistic = 37.189, p < 0.001$).

In addition, the correlations between the speech produced by the wearers and the key variables (MADRS=Montgomery-Asberg Depression Rating Scale; APS=Attention and Psychomotor Speed; DSSI=Duke Social Support Index; LSNS-R=Lubben Social Network Scale-Revised) for participants with Late-Life Depression (LLD; N=29) and healthy controls

(N=29) are also presented in the Figs. 4.22, 4.23, 4.24 and 4.25. The correlation figures are generated by Seaborn python library. Since LLD patients and control subjects differed so markedly in the quantity of speech they encountered and the quantity of speech that they produced, it is however surprising that the wearer speech did not correlate with the clinical scales of depression in the LLD group. Similarly, since we considered the quantity of speech detected as a measure of social interaction, it is also unexpected that the wearer speech did not correlate with the self-report scales of social functioning. It could be that our measures of speech may reflect a more accurate measure of social interaction than the self-report scales, which are influenced by bias. Indeed, previous research has highlighted that a discrepancy between subjective and objective measures of social functioning may be due to a bias towards pessimism in participants with depression [142]. These results could also be explained by a floor effect in the speech data of the LLD group: there may have been insufficient variation to produce significant correlations. It is also possible that these measures of speech represent a depression-related construct that is independent of any of the other variables measured and that is not included in either depression scale. Another consideration is whether the speech measures reflect current depression (i.e. current 'state'), or whether they reflect something that distinguishes those who are prone to depression from those who are not (i.e. depressive 'trait'). Previous research suggests that changes in some aspects speech patterns have been found to be related to changes in depressed state in participants with depression, while others are related to a depressive trait [115]. If our speech measures reflect a trait of LLD, this may explain why speech did not correlate with MADRS or GDS-15, which measure depressive state.

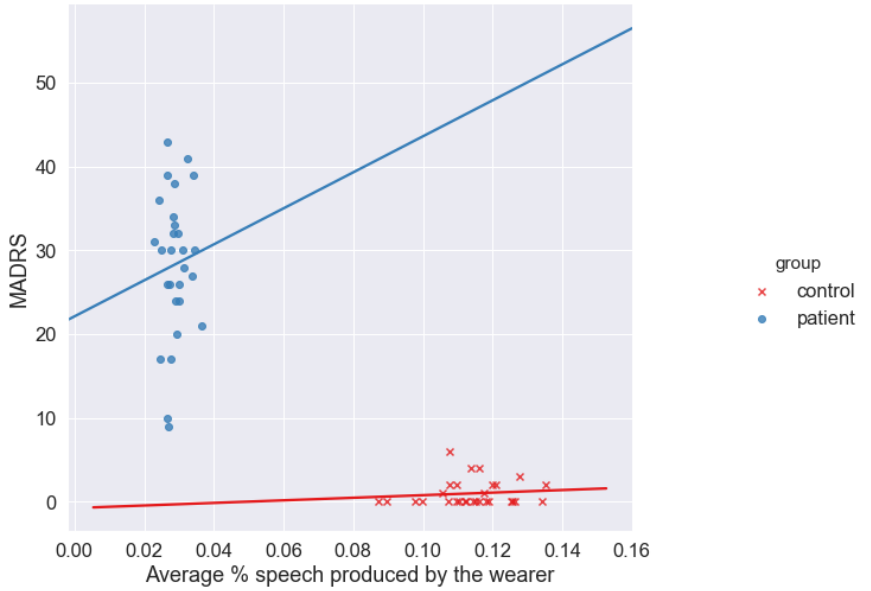

Fig. 4.22 Correlations between MADRS and mean percentage of speech produced by the wearer (out of all speech detected), for participants with Late-Life Depression (LLD; N=29) and healthy controls (N=29).

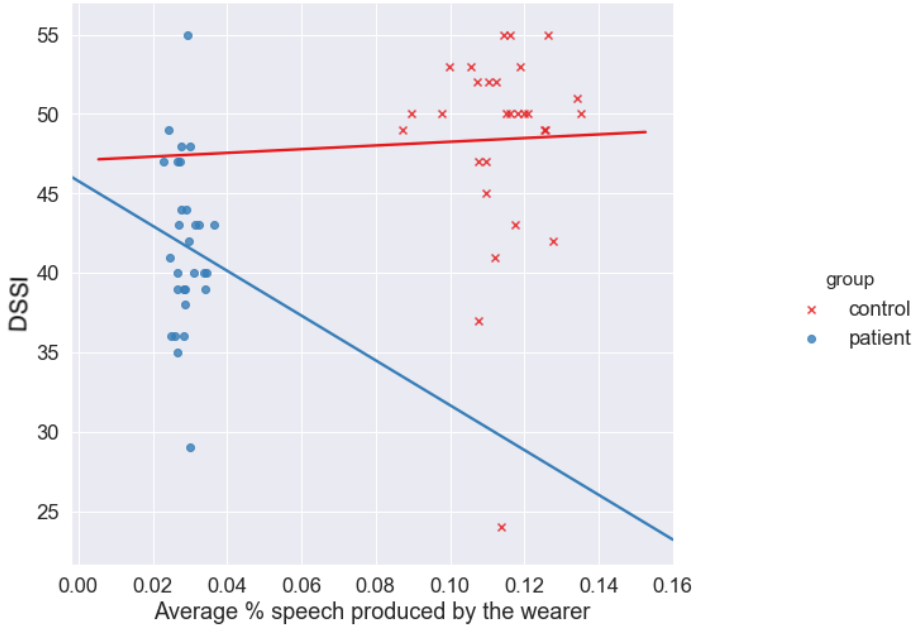

Fig. 4.23 Correlations between DSSI and mean percentage of speech produced by the wearer (out of all speech detected), for participants with Late-Life Depression (LLD; N=29) and healthy controls (N=29).

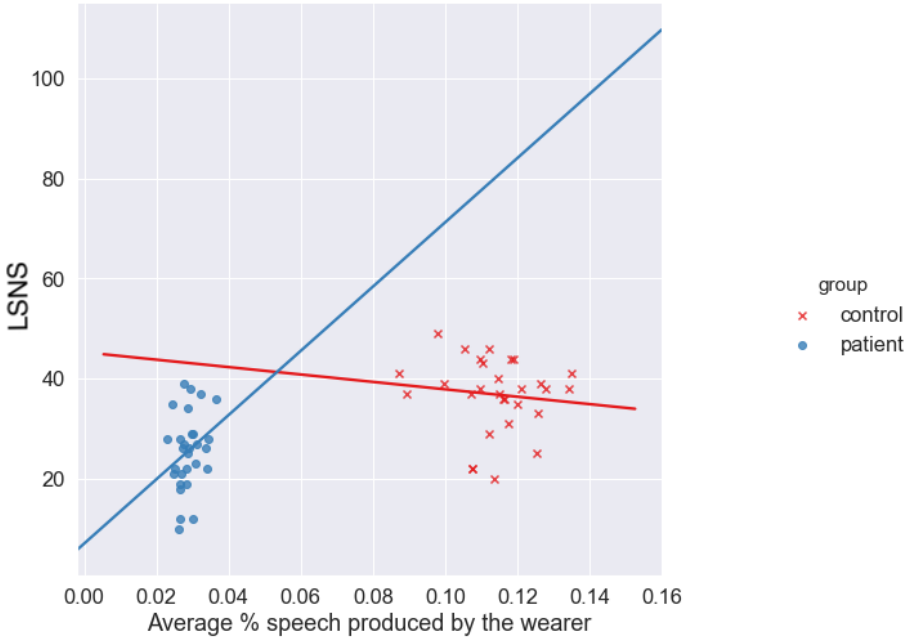

Fig. 4.24 Correlations between LSNS and mean percentage of speech produced by the wearer (out of all speech detected), for participants with Late-Life Depression (LLD; N=29) and healthy controls (N=29).

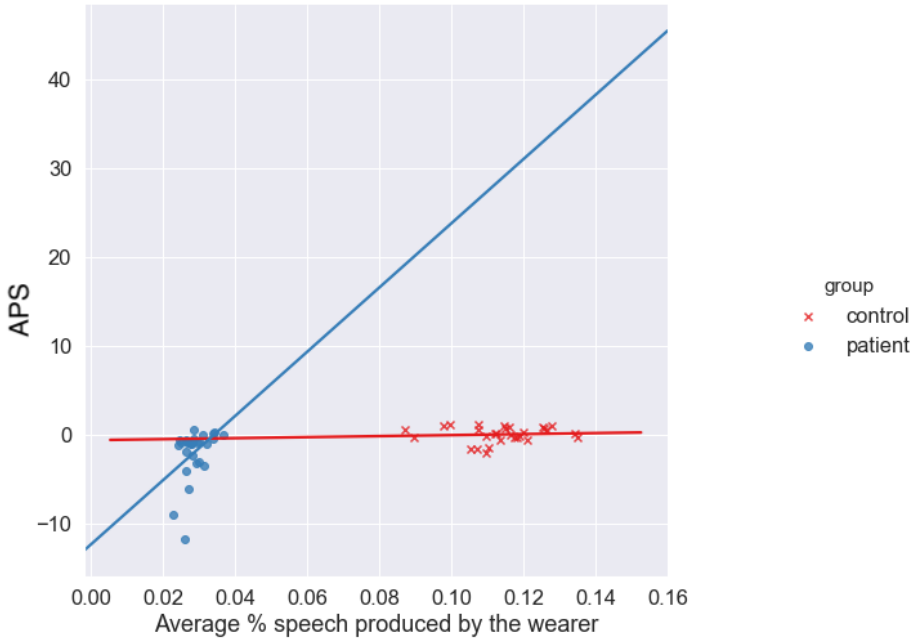

Fig. 4.25 Correlations between APS and mean percentage of speech produced by the wearer (out of all speech detected), for participants with Late-Life Depression (LLD; N=29) and healthy controls (N=29).

Finally, the wearer's speech measure may vary in accuracy for different cohorts, due to changes in the way people socialise and communicate (i.e. verbally versus non-verbally via technology). Nevertheless, the measures of speech presented here appear to be associated with depression and can accurately distinguish from controls, and may therefore be a useful marker for LLD. A particular strength of this work was that the device was unobtrusive and we found high adherence with wearing the device, demonstrating the feasibility of using such devices with older participants. If developed further, this measure has the potential to be used in screening for LLD, facilitating early diagnosis, and has implications for monitoring long-term health and recovery. The methods presented here provide a starting point for further research using raw sensor recordings and automatic analysis to investigate speech and social functioning in LLD.

4.4 Conclusion

In this chapter, I present a model for detecting the main wearer of the WAM device after the speech is detected by NatSpeech. This model is called NatWearer which based on CAE and BLSTM with Maxout layers. It has been designed to better detect the main wearer from the conversation speech predicted by NatSpeech. CAE is shown to have a substantial effect on the performance of learning the speech patterns of the main speaker and help the model achieve better performance in terms of F1-score in addition to PR-curves. Also, the performance of NatSpeech to detect the wearer is presented and shows a good F1-score. However, the PR-curves has a difficulty in some sessions which justify the reason of building new model. The novelty is the combination of CAE, BLSTM and Maxout layers with a combination of two optimisers (Adam and RMSPROP). The model is trained on the in-house dataset (Discussion dataset) and then applied to the Depression dataset and is capable of discriminating the depressed cohort from the control subjects.

Chapter 5

Physical Activity Analysis for people with Depression

The motivation for this chapter is analyse the physical activities for people with depression and this is the second aspect of the social interaction as stated in Chapter 1 (Introduction) and this is the third research question. Also this chapter is to give a conclusion for the previous speech measure and their correlation with the physical activities in order to answer the fourth research question and give a complete story for the social interactions analysis.

5.1 Introduction

Physical activity and sleep patterns have a substantial health implications [18],[149]. The analysis of daily life activities of observed people provides patterns of specific activities and assists them to relieve from a problem. So, people who are sitting for longer times, that may lead to cognition problems or having dementia at some point or having psychiatric problem and insomnia if they sleep for longer times during the day time and have reduced amount of time at night. However, regular exercise, organised and controlled eating healthy food and other structure daily life activities may lead the person to healthier life routine [175]. Daily life activities in addition can help to alleviate the problem of sudden fall for elderly people. The daily activities are intrinsically classified into 3 groups [124]: temporal activities, basic activities and complex activities. The temporal activities are the temporal movement patterns that are changing positions over time. Examples of temporal activities includes sit-to-stand, push-ups, pull-ups and sit-to-lie [135]. The basic activities are the activities that have longer duration than the temporal ones. These activities include sleeping, walking, running, cycling,

cooking, using vehicle and lying [62]. Complex activities are synchronous or consecutive forms of basic activities such as smoking, coffee time, etc. [175].

Physical activity (or movement patterns) is associated with late-life depression (LLD) bi-directionally. It is proven clinically that LLD is correlated with a decline in daily life activities [121]. This relationship has lead to measuring the daily activities in the real world on an ongoing basis especially for those with Depression. The issue is that LLD patients are assessed by psychiatrists using self reports, surveys and questionnaires which are not sufficient for accurate measure of the social interactions [121]. These self reports, surveys and questionnaires are usually biased and subject to measurement error [29]. There are various analysis like stated in Chapter 2 for the use of actigraphy to assess daily activities. Notwithstanding, this analysis lacks the interpretability of light, moderate and vigorous tasks especially in elderly people [64],[84].

Considering the fast improvements physical activity analysis has achieved, it has been recommended to use the raw accelerometer data (3D acceleration signals) rather than aggregating epoch level features extracted from the raw signal [138]. Unlike actigraphy, WAM device (discussed in Chapter 3) allows for direct sensor reading of the 3D signal and hence the analysis can be applied to the raw data. This chapter presents objective measures of physical activity behaviours where a pretrained model [177] (See section 2.2.3) is leveraged and used to predict physical activities (sedentary actions, walking, mixed activities (light and moderate tasks), cycling, using vehicles as well as sleep patterns) then generate high-level features from these predictions. This answer the third research question and the second aspect of the social interactions. The high-level features are processed in a variety of ways through a hand-crafted feature engineering process to establish their capacity to discriminate between controls and patients in the depression dataset. This is done first feature by feature using standard statistical analysis and then in combination by training Random Forest models and evaluating with nested cross validation.

5.2 Method

The method begins with reading the data from the WAM device and converting it to a readable format (comma separated version or .CSV file format) that can be analysed. Then this data is preprocessed for feature extraction process and subsequently a pretrained model [177] is used predict each activity that every participant performed. Then the predicted activities are used to extract biomarkers followed by thorough statistical analysis for the biomarkers and their association with the Depression key variables. Finally, the predicted activities are exploited to create a high level features which are meaningful compared to the features

generated used the pretrained model. These high level features are fed to another random forest model trained and evaluated using nested cross validation.

5.2.1 Preprocessing

The WAM device supports a serial based API over a USB port and stores its data to an .OMX in addition to the audio files with .EWV format. WAM records data at the naturally slightly-variable rate of the underlying sensors. The audio data is recorded at 8kHz and the movement data is recorded at 50Hz. The .OMX file format contains the compressed raw sensor-data for the accelerometer, light and proximity sensors altogether on its internal storage. In order to extract the 3D accelerometer signal only, OpenMovement (OM) conversion tool [1] is exploited. The OM conversion tool is a C++ code that converts the input raw .OMX file format to .CSV file format. The timestamps for the conversion is up to one sample per sector timestamps which are very accurate. Furthermore, OM converter can estimate the non-wear time and interpolate the data as well. The interpolation of data can be achieved by any of the three methods: nearest-neighbour, linear interpolation and cubic interpolation. In this thesis, cubic interpolation is used. Moreover, OM converter has the capability to create signal vector magnitude in .CSV file format by setting the epoch length of the data frame in seconds. The epoch length is the length of time each signal vector magnitude averaging is made over. The signal vector magnitude used in OM converter is given by Eq. 5.1. In this thesis, the .OMX format is converted to .CSV file format which contains the columns of real timestamp, x, y, z coordinates respectively. Subsequently, this data is split into frames of non-overlapping 30 sec window length like in [37]. The frames are then ready for feature extraction the features.

$$\sum |\sqrt{x^2+y^2+z^2}-1| \tag{5.1}$$

5.2.2 Feature Extraction

After the frames are created, a vector of 130-dimensional features is calculated. The choice of these features is based on a combination of time-domain and frequency-domain features as in the state-of-art studies [104],[62],[104],[160],[175],[173],[3]. The list of features are illustrated in Table 5.1. Subsequently, the features are ready to be fed to the classifier. The classifier in this case is random forest and hidden Markov models is used to handle the temporal dependence as discussed in Chapter 2.

Due to the absence of the annotation for the Depression dataset which is privacy sensitive, the pretrained model of random forest is utilised. This pre-trained model is adopting HMM

[1] https://github.com/digitalinteraction/openmovement/tree/master/Software/AX3/omconvert

Time-domain features	Frequency-domain features
euclidean norm minus 1 *	Fast Fourier Transform coefficients 1–15 Hz
mean	frequency bands
standard deviation	power bands
median	mean power deviation
minimum	roll
maximum	pitch
25th percentiles	yaw
75th percentile	
coefficient of variation	
mean amplitude deviation	
skew	
kurtosis	
x, y, z coordinates correlations	

Table 5.1 List of time-domain features as well as frequency domain features. These are the features used to be fed to the classifier for predicting the activity. * euclidean norm is calculated and then all negative values are set to 0.

to handle the temporal variation and is trained over 132 adults with data recorded in realistic free-living style that created naturalistic behaviours [177],[38]. This pre-trained model is available at https://github.com/activityMonitoring/biobankAccelerometerAnalysis. The input to the pre-trained model is the features listed in Table 5.1. Then the predictions for each activity is obtained. The activity predictions are used to extract meaningful biomarkers and are then utilised to extract high-level that gives better representation and interpretation of the activities. These features include the number of bouts for each activity, mean of the bout duration of each activity, standard deviation of bout duration for each activity, percentage of each activity. This combination of hand-crafted features that are fed to train RF models are utilised as biomarker as well.

The reasons to utilise RF with the high-level features are: 1) the attributes (features vectors) are not required to be normalised like in any other classification algorithm and thus this reduces the burden of pre-processing operations that need to performed in order to make the classifier work better; 2) The RF algorithm does not use the whole features in each tree but effectively do selection and therefore no need to use any feature selection algorithm to solve this issue; 3) The growth of the decision trees to their maximum depth is performed separately for each of them utilising the randomly selected set of features and hence this make the training dataset not to overfit; and 4) the randomness in selecting both of the set of instances and the set of features will ensure that there is a low correlation between trees and especially for CART due to its deterministic nature [179].

5.3 Results and Discussion

After the features are extracted as discussed in the previous section, these features will be fed to the pretrained RF-HMM to get predictions for each activity on the unannotated Depression Dataset. The predicted activities from the model are sleep, sedentary, cycling, using vehicles, walking or mixed activities. The physical activity data is recorded for 29 control subjects and 29 depressed patients. In the dataset, there is no self-reporting on the physical activities performed during the week of recording. Thirteen selected activity predictions for control subjects and depressed cohort that have interesting patterns are shown in Figs. 5.1 through Fig. 5.13. The plots are produced by the plotting tool in https://github.com/activityMonitoring/biobankAccelerometerAnalysis.

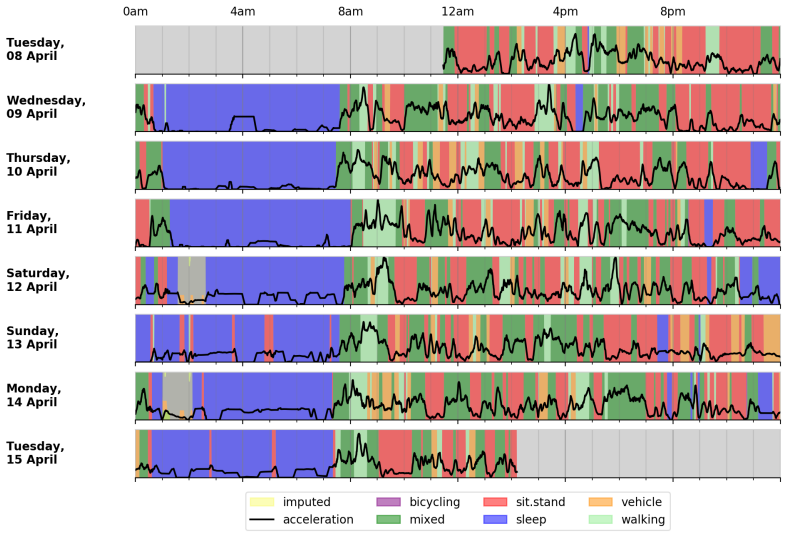

Fig. 5.1 Physical activity predictions for participant ID=1112 from the control group.

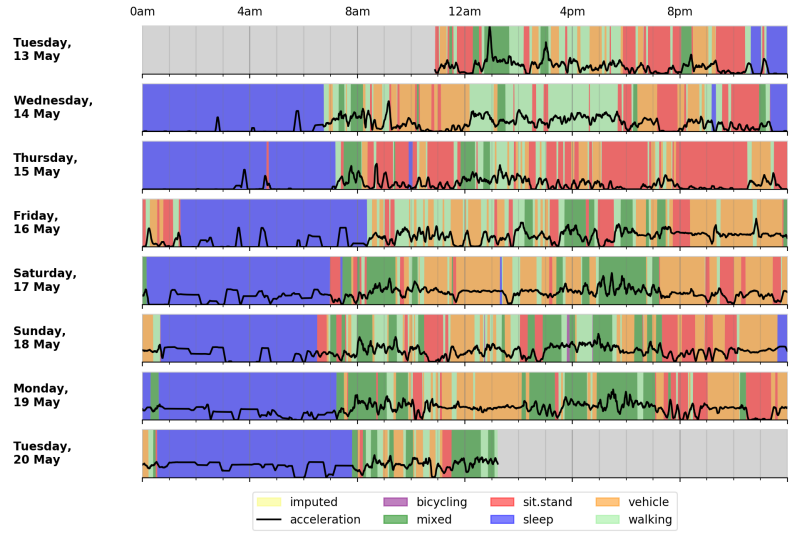

Fig. 5.2 Physical activity predictions for participant ID=1589 from the control group.

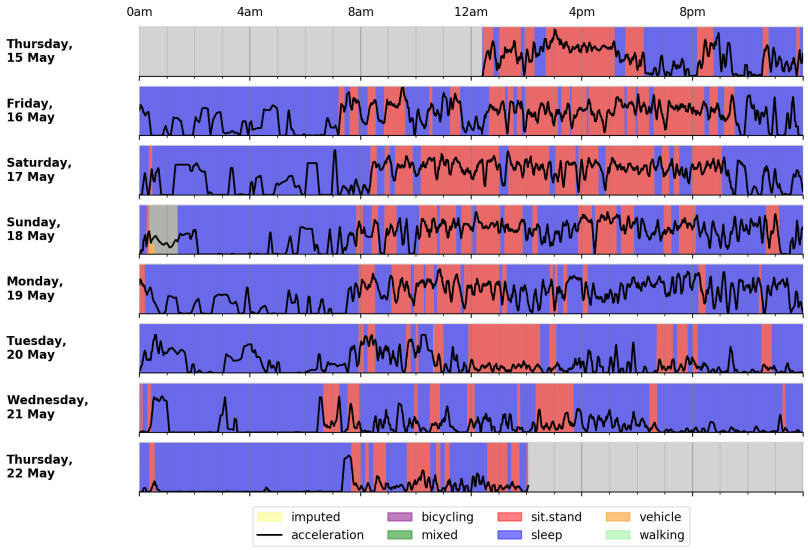

Fig. 5.3 Physical activity predictions for participant ID=6549 from the control group.

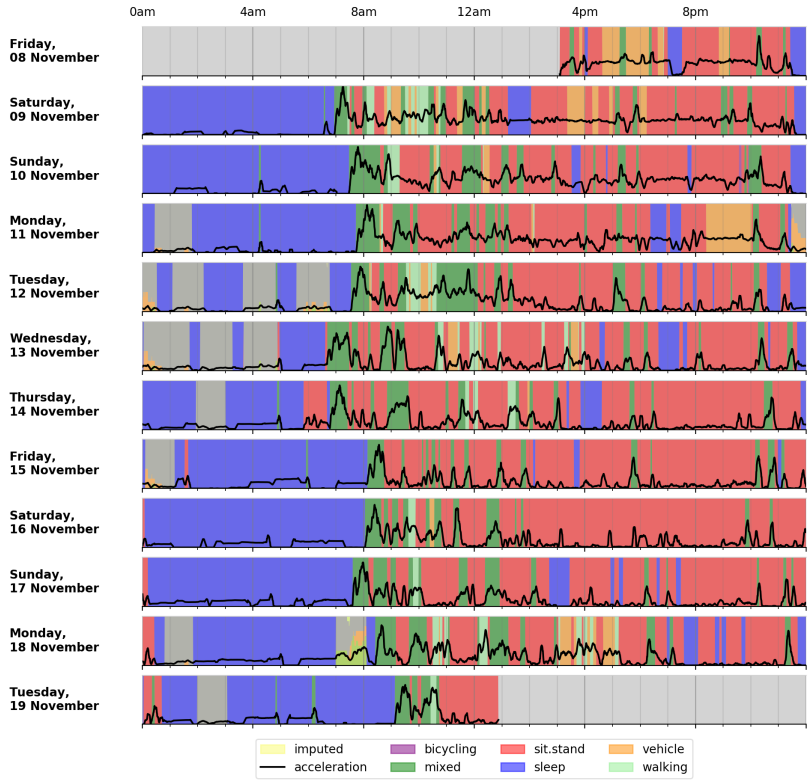

Fig. 5.4 Physical activity predictions for participant ID=4938 from the control group.

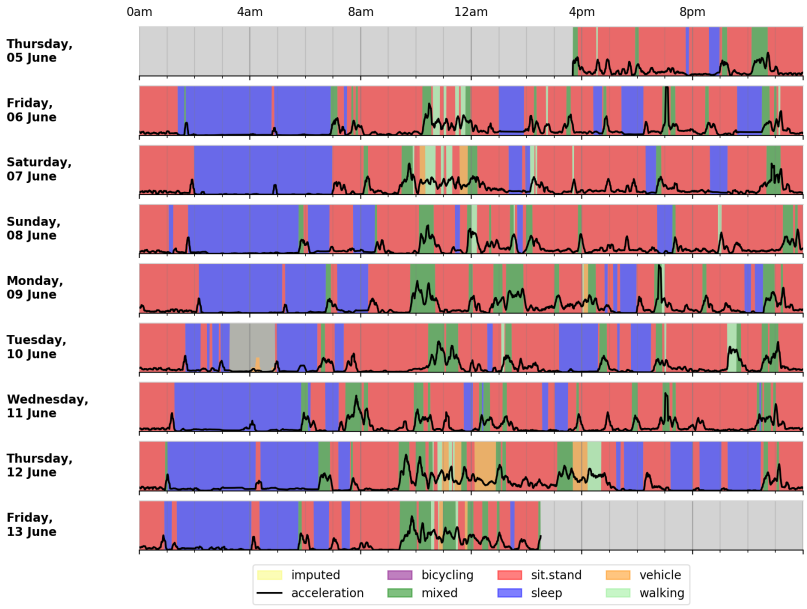

Fig. 5.5 Physical activity predictions for participant ID=6548 from the control group.

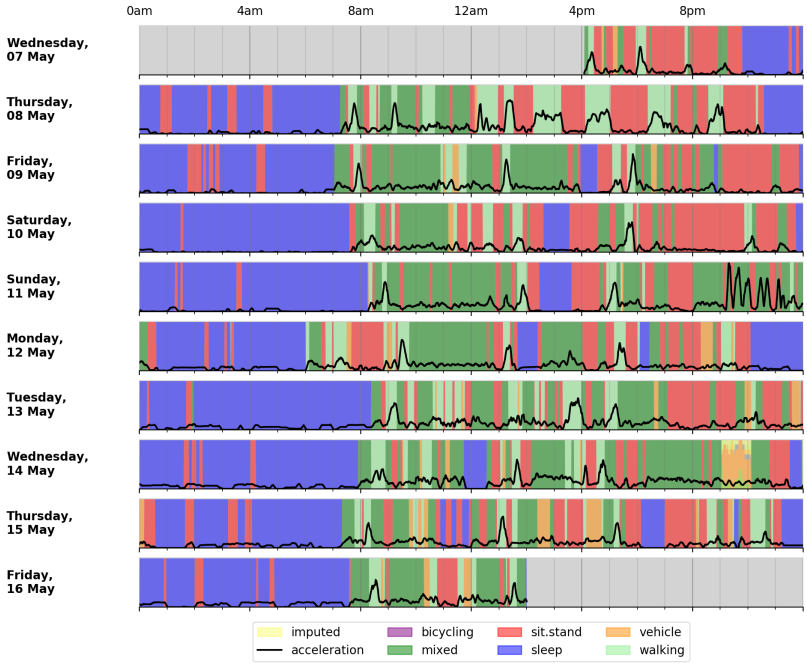

Fig. 5.6 Physical activity predictions for participant ID=6611 from the control group.

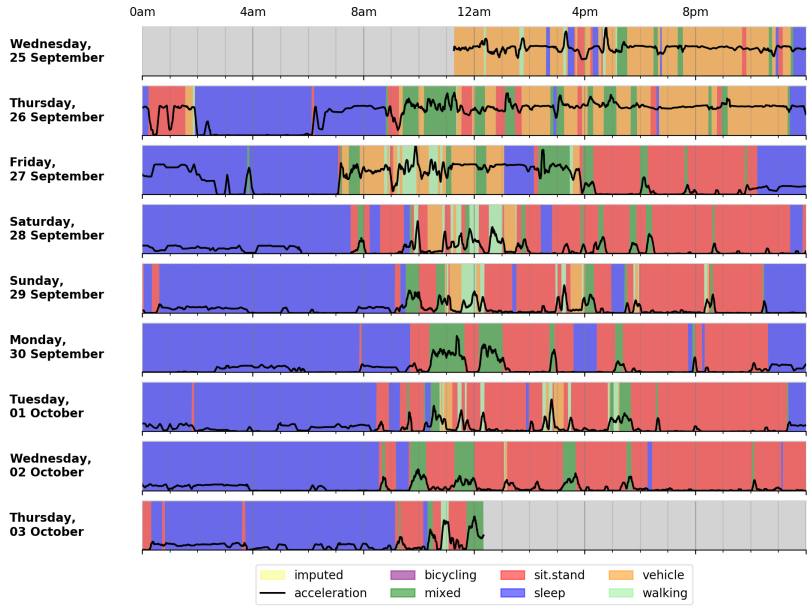

Fig. 5.7 Physical activity predictions for participant ID=1152 from the patient group.

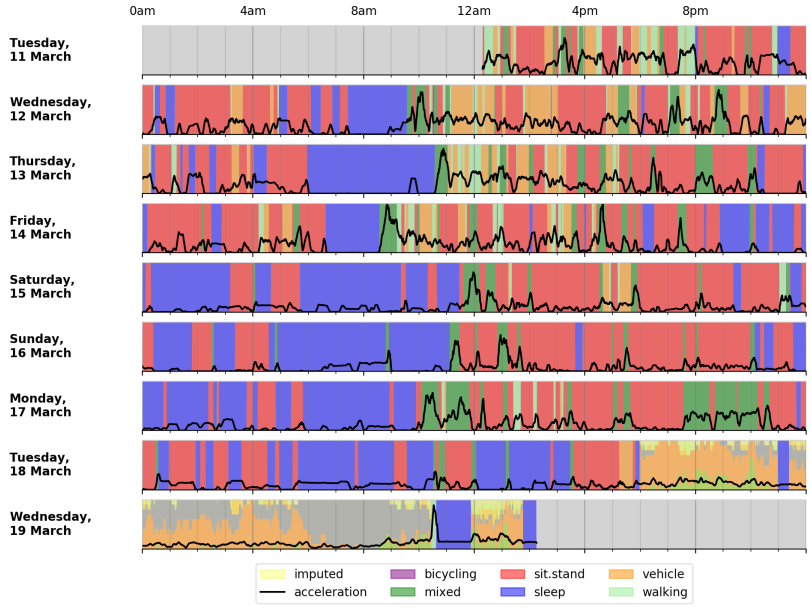

Fig. 5.8 Physical activity predictions for participant ID=2151 from the patient group.

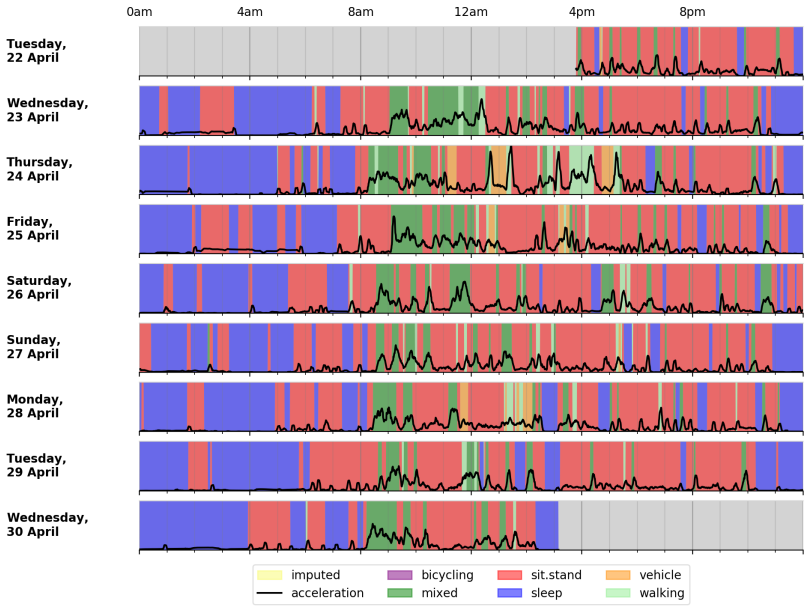

Fig. 5.9 Physical activity predictions for participant ID=3586 from the patient group.

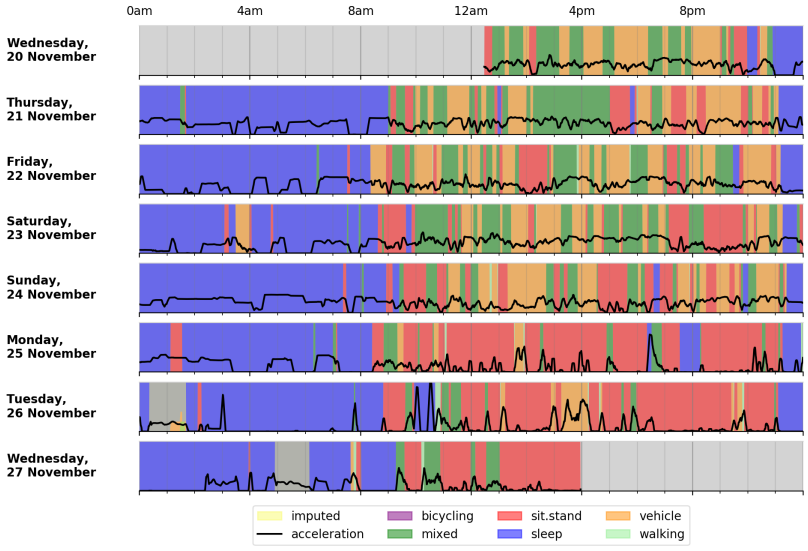

Fig. 5.10 Physical activity predictions for participant ID=2449 from the patient group.

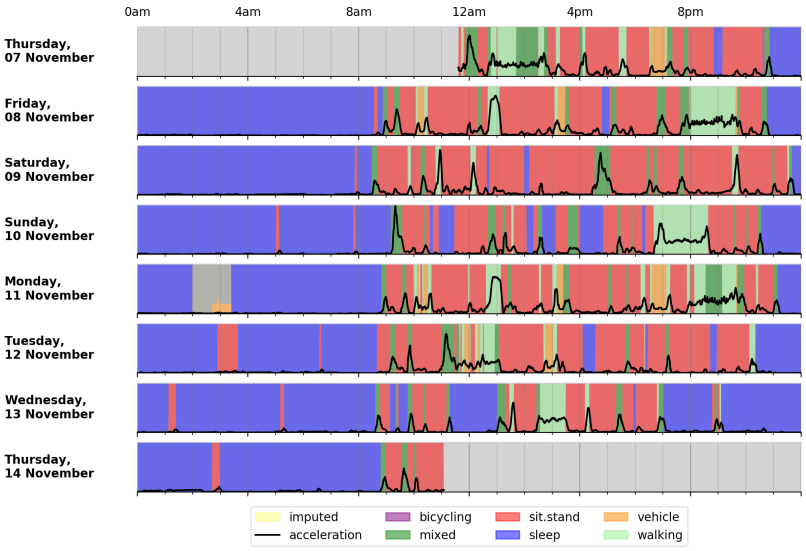

Fig. 5.11 Physical activity predictions for participant ID=2834 from the patient group.

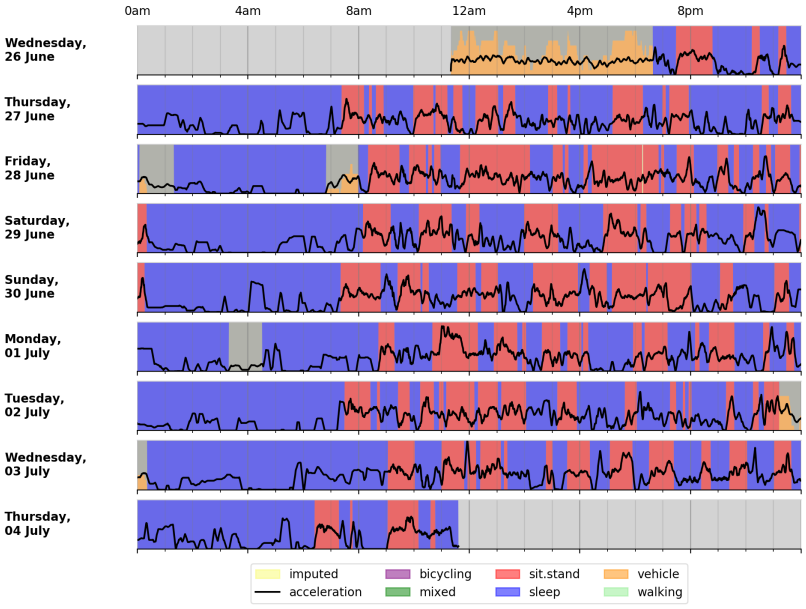

Fig. 5.12 Physical activity predictions for participant ID=3168 from the patient group.

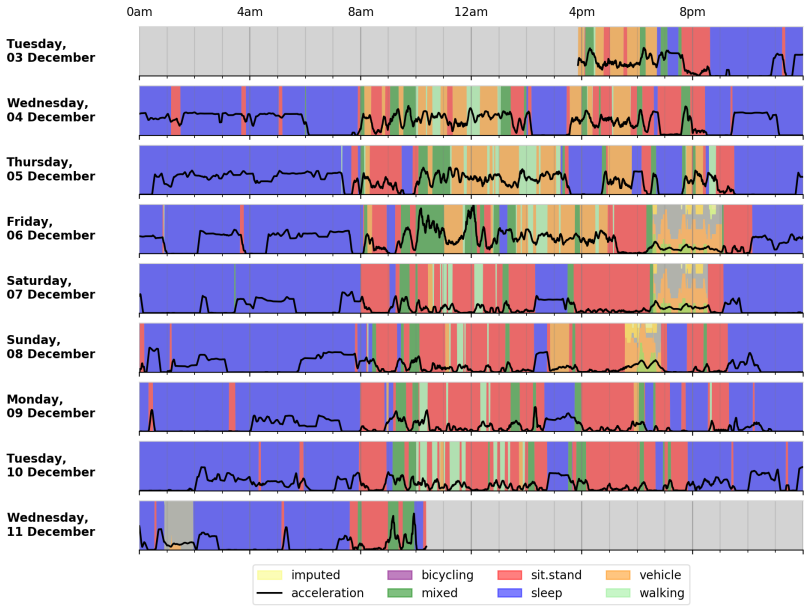

Fig. 5.13 Physical activity predictions for participant ID=3762 from the patient group.

The presented figures have some interesting patterns for each activity. In the figures the grey pattern within any activity predictions represents missing data from the device due to the system restart or a fault happen but this happened rarely in all of the recordings. The control subjects have normal sleep patterns and this is according the number of minutes and a variety of activities such as walking, mixed activities especially in Fig. 5.1. For participant ID=1589 in Fig. 5.2, the activities predictions show regular walking and mixed activities during the day more than any other participant. However, there is a large amount of the vehicle usage patterns and also the fourth and fifth days show late sleep. Large quantity of sleep patterns have been shown in Fig. 5.3 where most of the day basically sleeping and sedentary actions. Since there is no self-reports for the activities, there is no evidence whether the participant was sick or that might depend on the time of the year affected by the weather. This figure represents an outlier in the control group especially the sixth day where the majority of the times during the day is just sleep. In participant ID=4938 in Fig. 5.4, there are a lot of sedentary actions within the day although in between there is some walking, mixed activities or using vehicle. In terms of abnormal sleep or sleep interruptions along with a lot of sedentary actions, participant ID=6548 is considered to have such kinds of patterns. Apart form these special cases, the overall predictions of other participants are normal in terms of either sleep or other daily life activities.

On the other hand, in the depressed cohort, the majority of participants are having less in the walking and mixed activities compared to the healthy control subjects and sometimes more sleep time. An example from the cases that show interruption of sleep are displayed in Fig. 5.8. One of the depressed cohort which shows sleeping most of the day and have sedentary actions in between which is more than any of the other participants even the one with similar patters in the healthy control group. A lot of sedentary actions with a little bit of either walking or mixed activities is shown in Fig. 5.11 whilst the patient with ID=3586 (Fig 5.9) has also interruption in sleep where there are some sedentary actions are predicted with the night and also in Fig. 5.7 which shows some vehicle usage patterns. A variety of activities but with more sleeping can be seen in Fig. 5.10. The more from all of these prediction that in the depressed cohort, there is more sleep, more sedentary actions and less in both walking and mixed activities.

As an overall trend, in both groups the cycling rarely happens and plenty of sedentary activities patterns. Although in most depressed patients the sleep patterns is more than normal, in some control subjects they also have the same case. Furthermore, sedentary actions are noticeable in both groups but does not seem to be more in one group than another. On contrary to these actions, walking ($mean$=14.7, std=8.1 for controls and $mean$=10.1, std=8 for patients), mixed activities ($mean$=26.8, std=16.1 for controls and $mean$=18, std=14.5 for patients) and vehicle usage ($mean$=25.9, std=15.6 for controls and $mean$=20, std=14.6 for patients) is apparent in healthy control subjects rather depressed cohort. The mixed activities term is defined as a mixture of light and moderate tasks. In order to investigate the distribution of the activities in the two groups, normality test is performed first. The hypothesis is based on the clinically proved fact that there is a decline in physical activities in depressed patients [121]. The assumption is the average duration of walking, the average duration of mixed activities and the average duration of vehicle usage can be used as a biomarkers for Depression. To achieve this, the predicted epochs (30 sec) of each activity is counted and then compute the mean of the bouts of that activity. The bout is a periodic continuous activity. Deciding the statistical test is relying on whether the data is normally distributed or not. By using D'Agostino's K^2, the Figures 5.14, 5.15 and 5.16 represent the normality test of the average duration of walking, the average duration of mixed activities and the average duration of vehicle usage. The null hypothesis for this test is that the data is normally distributed. These figures show non-normal distribution for them. So, the suitable statistical test will be Mann-Whitney test. The test shows a p-value less than the (significance level=0.5) which means to reject the null hypothesis. Therefore, the data does not look Gaussian (normally distributed).

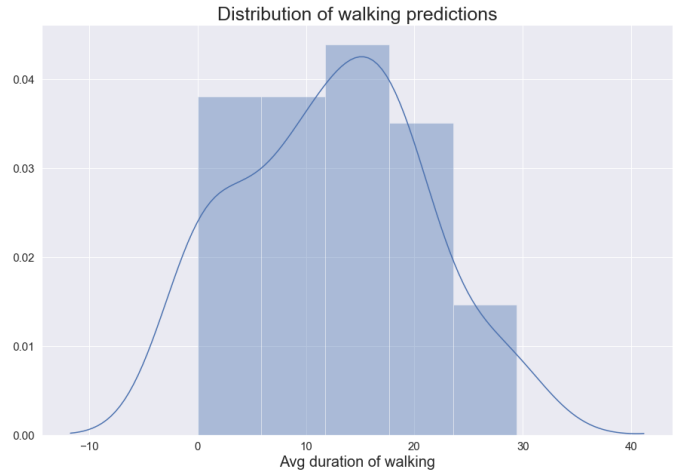

Fig. 5.14 The normality test of the average duration of walking predictions. According to D'Agostino's K^2 normality test (statistic=2.37, p-value=0.031), the data shows non-normal distribution.

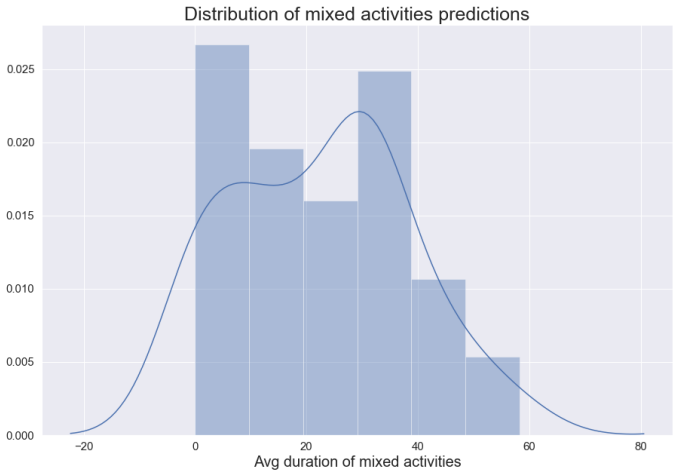

Fig. 5.15 The normality test of the average duration of mixed activities predictions. According to D'Agostino's K^2 normality test (statistic=3.09, p-value=0.04), the data shows non-normal distribution.

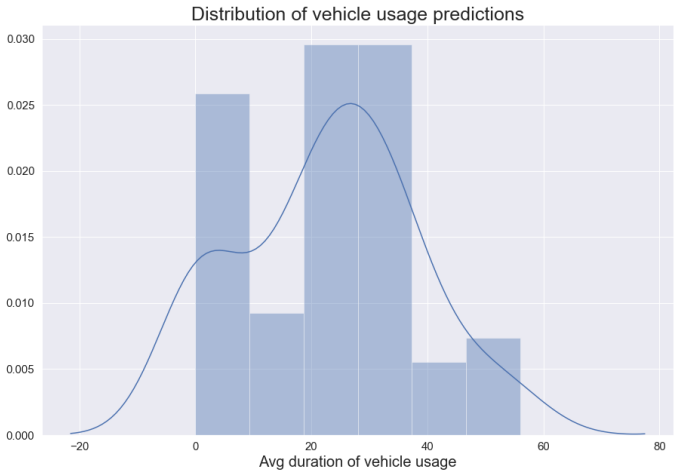

Fig. 5.16 The normality test of the average duration of vehicle usage predictions. According to D'Agostino's K^2 normality test (statistic=0.94, p-value=0.01), the data shows non-normal distribution.

Performing the Mann-Whitney test on the average duration of walking bouts, the test gives (statistic=265, $p-value$=0.008 $<$ 0.01)) which shows a highly significant difference between the healthy control subjects and the depressed patients. Likewise, running the same test on the average duration of mixed activities bouts and average duration of vehicle usage, the test gives (statistic=287.0, $p-value$=0.019 $<$ 0.05) and (statistic=295.0, $p-value$=0.019 $<$ 0.05) for mixed activities and vehicle usage respectively which shows also a significant difference between the healthy controls and the depressed patients. These biomarkers show that the two groups differed in the average duration of walking and performing mixed activities. This is also shown in Figs. 5.17, 5.18 and 5.19. The non-significant variables are also shown in Table 5.2. The main reason for the non-significance in both groups is that the cycling rarely happens. This is clear since the participant are elderly people. The sleep and sedentary patterns occur a lot in both of the groups which lead to no difference between the groups.

Table 5.2 The non-significant physical activity markers for LLD.

Activity	Statistic	p-value
Cycling	377.5	0.23
Sedentary	389.0	0.31
Sleep	340.0	0.11

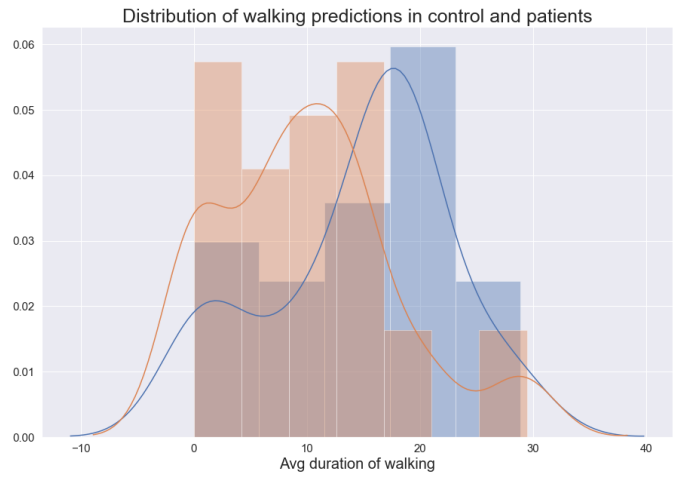

Fig. 5.17 The distribution of mixed activities predictions for the two groups. The blue histogram represents the control group while the orange represents the depressed cohort. The figure shows different distributions.

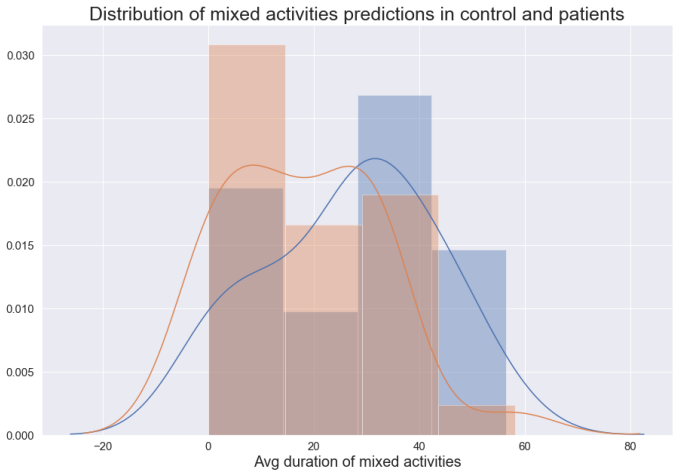

Fig. 5.18 The distribution of mixed activities predictions for the two groups. The blue histogram represents the control group while the orange represents the depressed cohort. The figure shows different distributions.

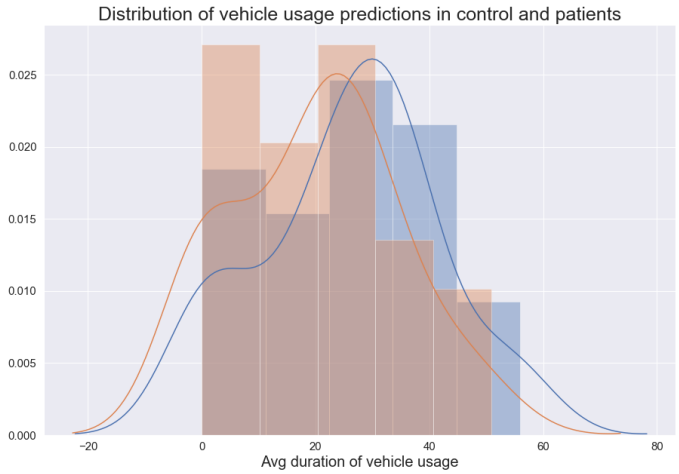

Fig. 5.19 The distribution of vehicle usage predictions for the two groups. The blue histogram represents the control group while the orange represents the depressed cohort. The figure shows different distributions.

Further analysis is done by looking over the correlation with the key clinical measures discussed in the previous chapters such as MADRS, APS, DSSI and LSNS-R. The correlations between the physical activities done by the wearer especially walking and mixed activities are presented in Figs. 5.20 through to 5.31. These figures show significant difference between the groups and the clinical variables for participants with Late-Life Depression (LLD; N=29) and healthy controls (N=29). The predictions figures discussed before displayed a substantial difference in the quantity of bouts predicted and also and the length of these bouts. However, Only vehicle usage within the three markers is correlating with MADRS and not the other clinical scales of depression in the LLD group. The correlation with MADRS in case of vehicle usage is significant positive correlation ($corr$=0.44, $p-value$=0.02). The reason for that could be the bias that the self-report scales has which is often affected by the mood [142]. Another reason that there may have been insufficient variation to produce significant correlations due to the small number of participants. It is also possible that these measures of physical activities represent a depression-related construct that is independent of any of the other variables measured and that is not included in either depression scale.

The intrinsic point for the social interactions analysis is not only correlating the movement markers with the key clinical variable but also to correlate with the speech variables (the total percentage of speech as well as the speech produced by the wearer) developed in this thesis. The correlation between average duration of walking and the percentage of speech is presented in Fig. 5.32. Note that the percentage of speech includes the wearer (main speaker) as well as the other speaker talking to the wearer. Using Pearson correlation, the test gives

corr=0.43, $p-value$=0.01 for the control group but gives *corr*=0.07, $p-value$=0.71 for the patients group. The figure and the test show the association between the average duration of walking and percentage of speech in the control group whilst in the patient group there is no association. The reason could be the fact that the depressed cohort is talking less most of the time and they might walk or not without talking. On the other hand, the correlation between the average duration of mixed activities and speech percentage is displayed in Fig. 5.33. The case is different here where there is a trend in the control group (*corr*=0.33, $p-value$=0.08) and no association at all in the patients group (*corr*=0.09, $p-value$=0.64). On contrary to the two markers: walking and mixed activity, there is no association at all between the third marker (vehicle usage) and speech. The test shows (*corr*=0.3, $p-value$=0.12) and (*corr*=0.22, $p-value$=0.28) for controls and patients respectively as shown in 5.34.

As regards the wearer's speech (speech produced by the WAM wearer), the correlation with the average duration of walking gives only significance association between them in case of healthy controls (*corr*=0.42, $p-value$=0.02) only like the case with the total speech but there is no correlation in case of depressed patients (*corr*=-0.04, $p-value$=0.85). This is also illustrated in Fig. 5.35. Furthermore, in case of the average duration of mixed activities and the wearer's speech, the correlation is shown in Fig. 5.36. Notwithstanding, this case does not show correlation with depressed patients (*corr*=0.09, $p-value$=0.64) or even with healthy controls (*corr*=0.3, $p-value$=0.10). Moreover, the correlation between the average duration of vehicle usage and wearer's speech is shown in 5.37. The test shows (*corr*=0.13, $p-value$=0.51) and (*corr*=0.08, $p-value$=0.68) for controls and patients respectively. This shows no correlation at all for both groups.

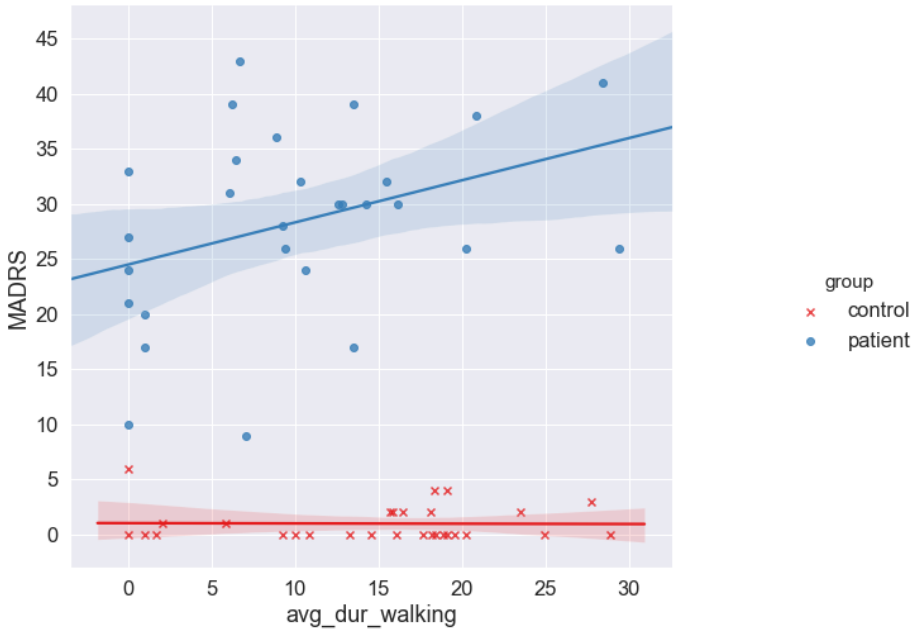

Fig. 5.20 The correlation between average duration of walking and MADRS for the control subjects (N=29) and the depressed patients (N=29).

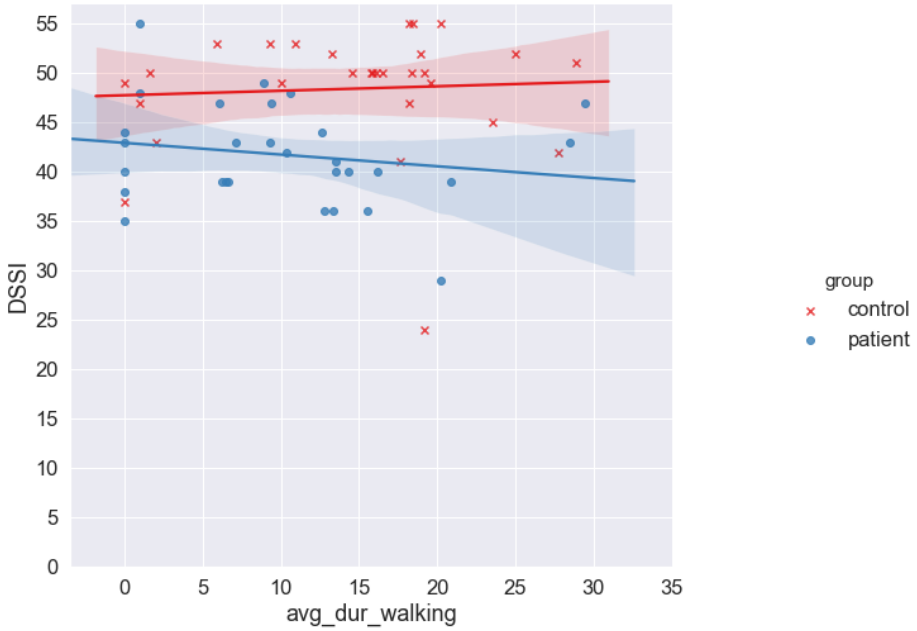

Fig. 5.21 The correlation between average duration of walking and DSSI for the control subjects (N=29) and the depressed patients (N=29)

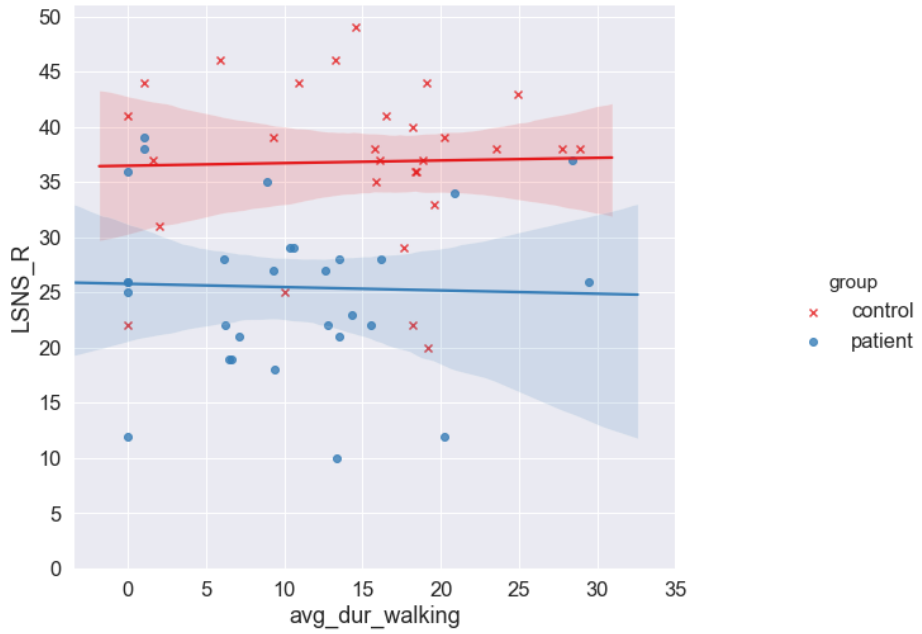

Fig. 5.22 The correlation between average duration of walking and LSNS for the control subjects (N=29) and the depressed patients (N=29)

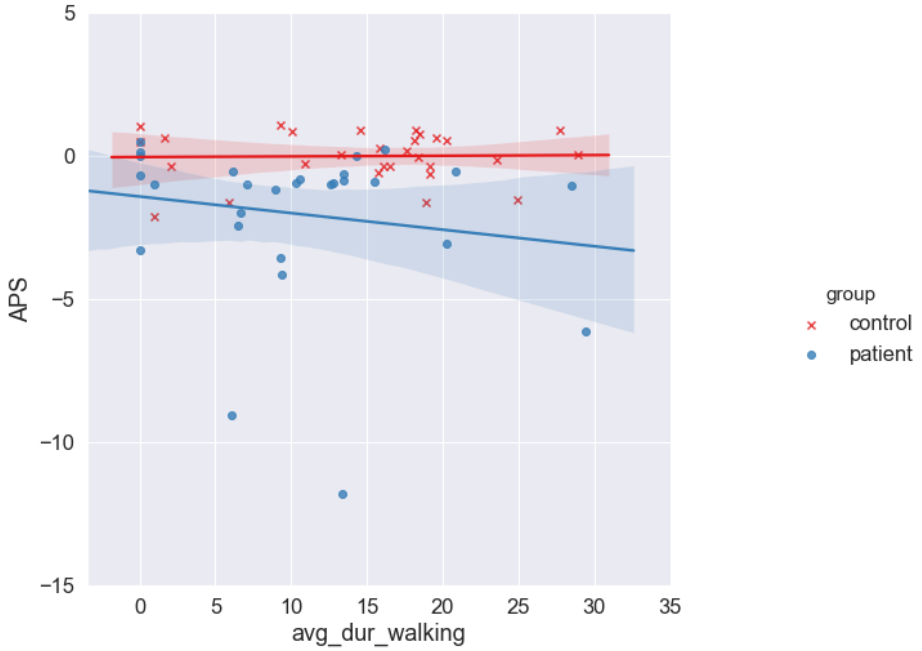

Fig. 5.23 The correlation between average duration of walking and APS for the control subjects (N=29) and the depressed patients (N=29)

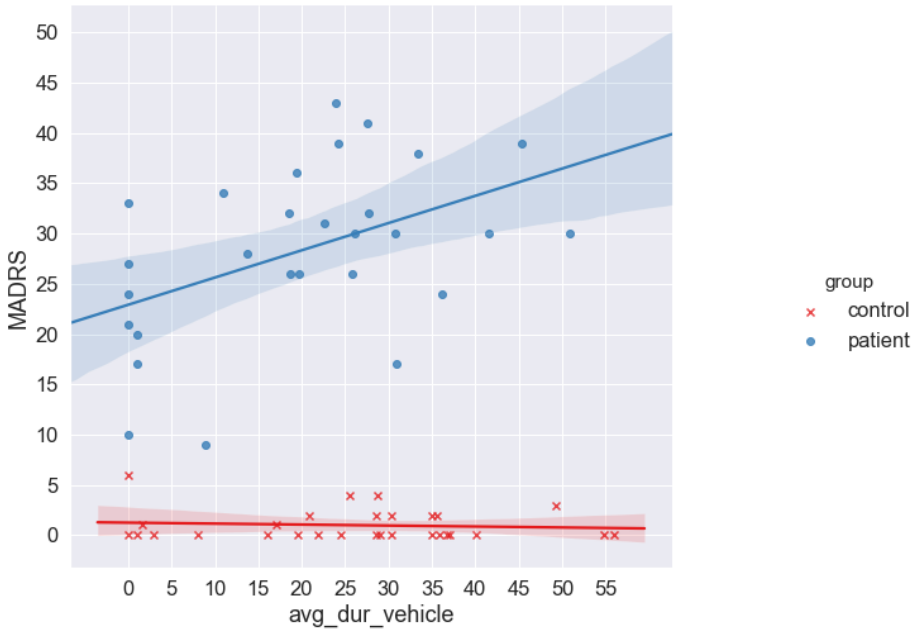

Fig. 5.24 The correlation between average duration of vehicle usage and MADRS for the control subjects (N=29) and the depressed patients (N=29).

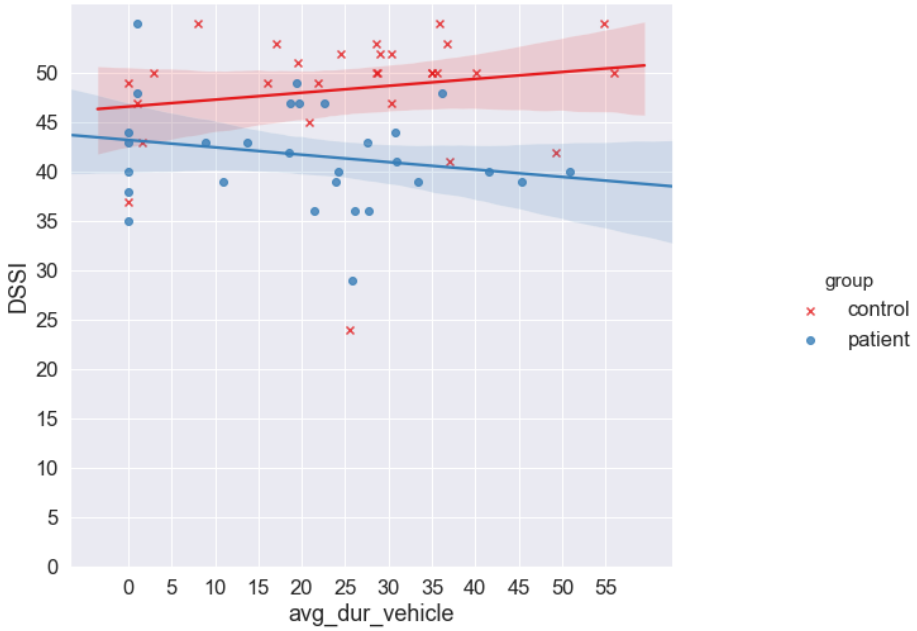

Fig. 5.25 The correlation between average duration of vehicle usage and DSSI for the control subjects (N=29) and the depressed patients (N=29)

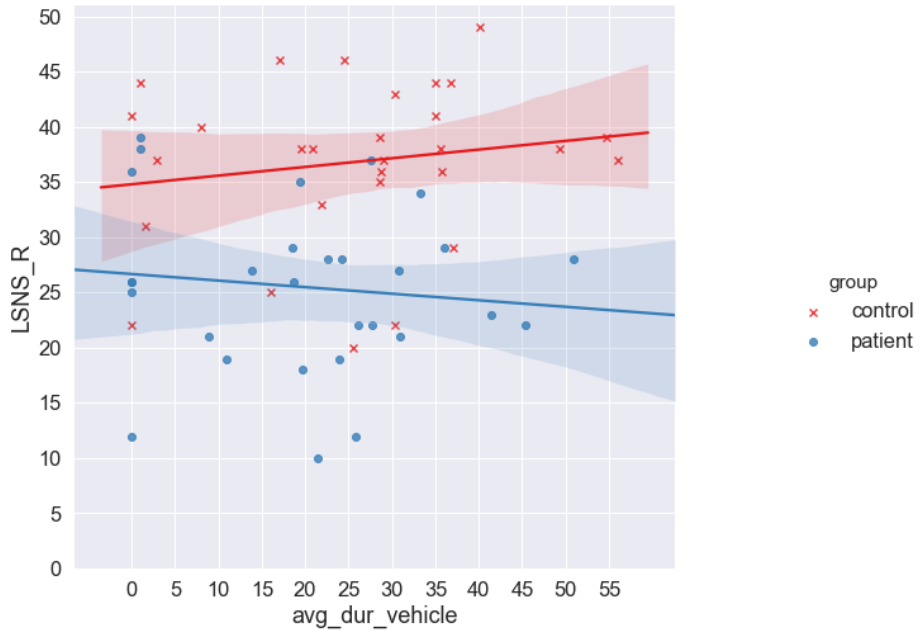

Fig. 5.26 The correlation between average duration of vehicle usage and LSNS for the control subjects (N=29) and the depressed patients (N=29)

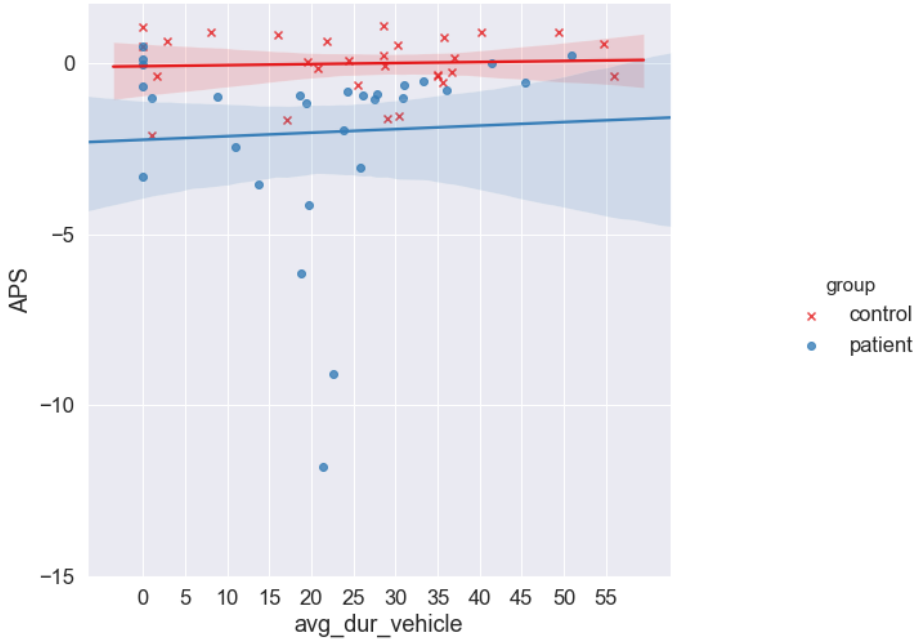

Fig. 5.27 The correlation between average duration of vehicle usage and APS for the control subjects (N=29) and the depressed patients (N=29)

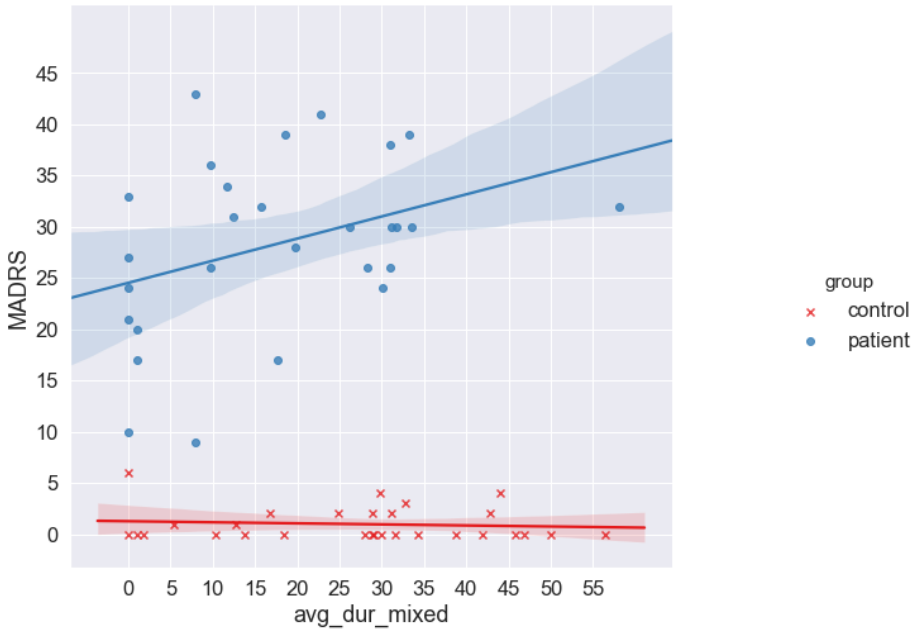

Fig. 5.28 The correlation between average duration of mixed activity and MADRS for the control subjects (N=29) and the depressed patients (N=29)

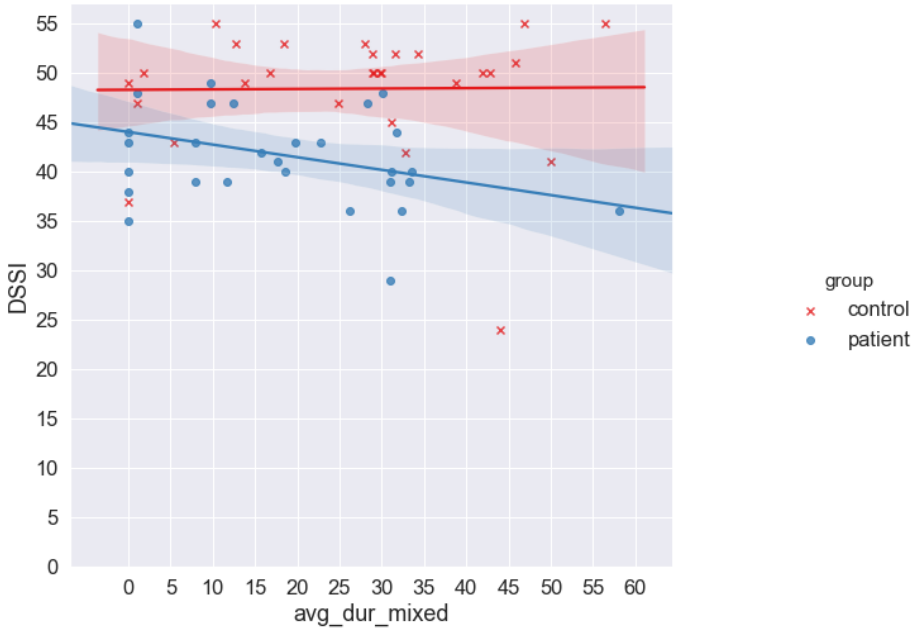

Fig. 5.29 The correlation between average duration of mixed activity and DSSI for the control subjects (N=29) and the depressed patients (N=29)

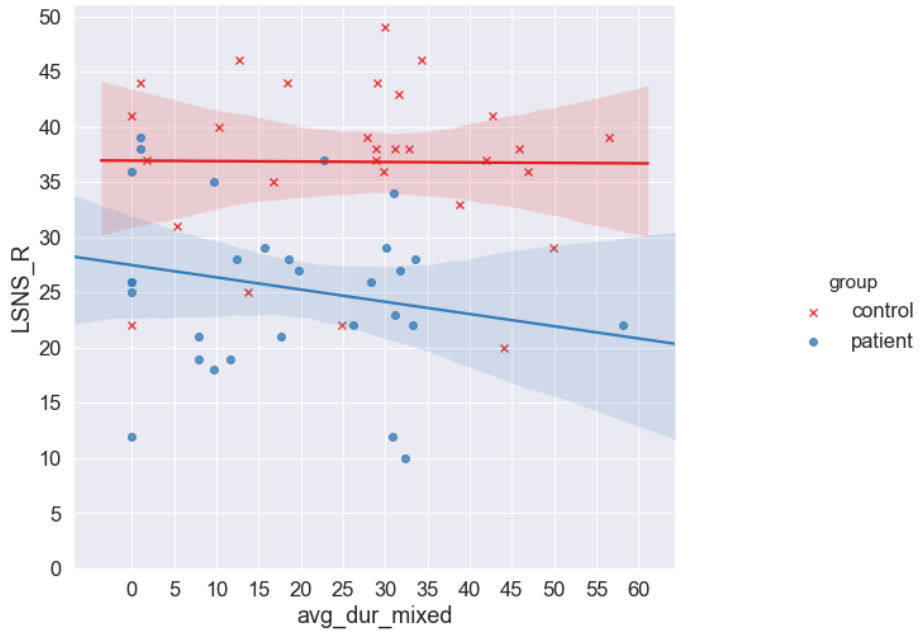

Fig. 5.30 The correlation between average duration of mixed activity and LSNS for the control subjects (N=29) and the depressed patients (N=29)

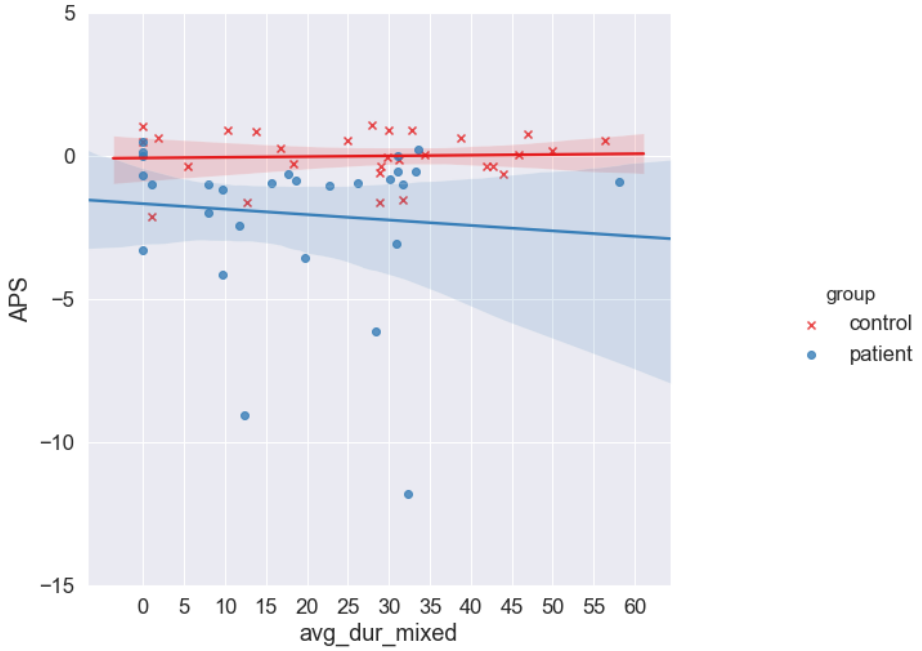

Fig. 5.31 The correlation between average duration of mixed activity and APS for the control subjects (N=29) and the depressed patients (N=29)

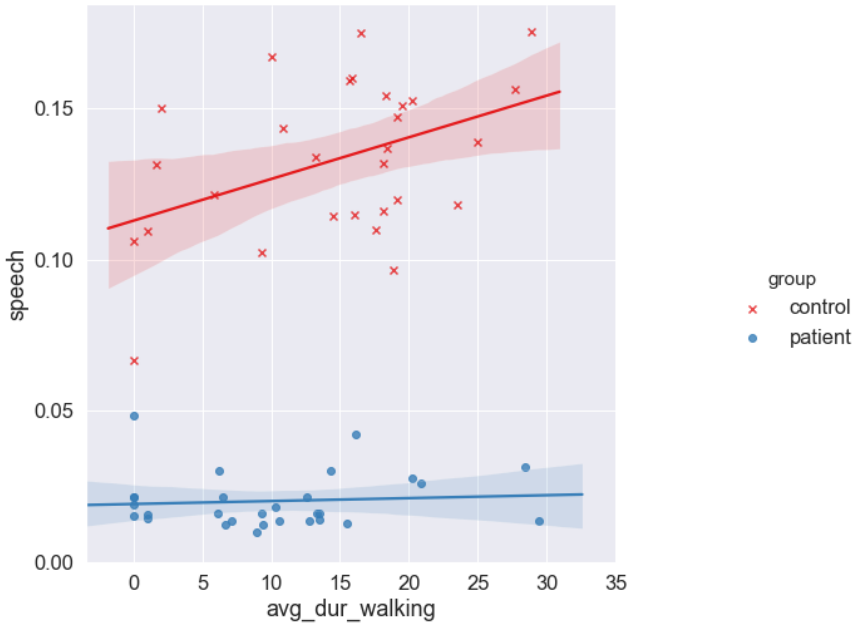

Fig. 5.32 The correlation between average duration of walking and total percentage of speech produced by the wearer and other speakers for the control subjects (N=29) and the depressed patients (N=29)

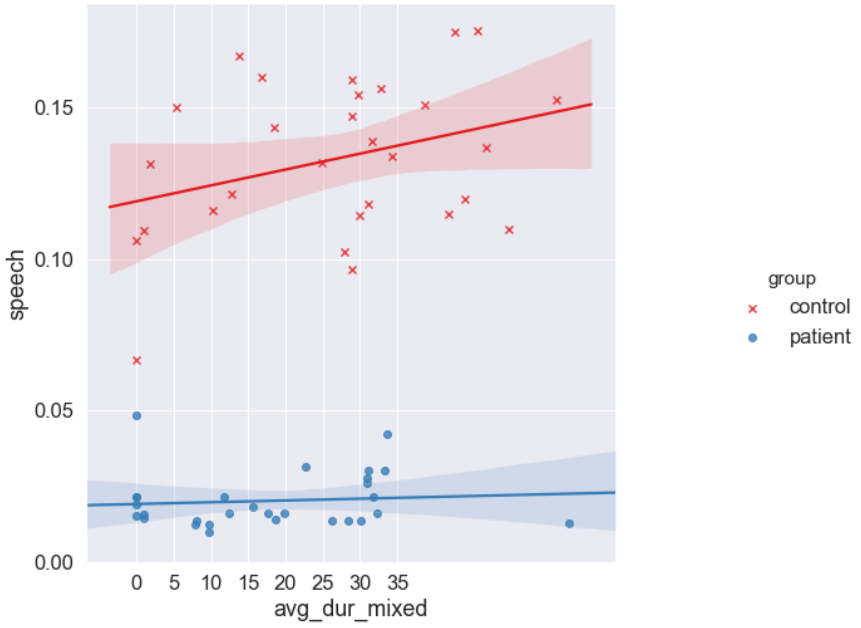

Fig. 5.33 The correlation between average duration of mixed activities and total percentage of speech produced by the wearer and other speakers for the control subjects (N=29) and the depressed patients (N=29)

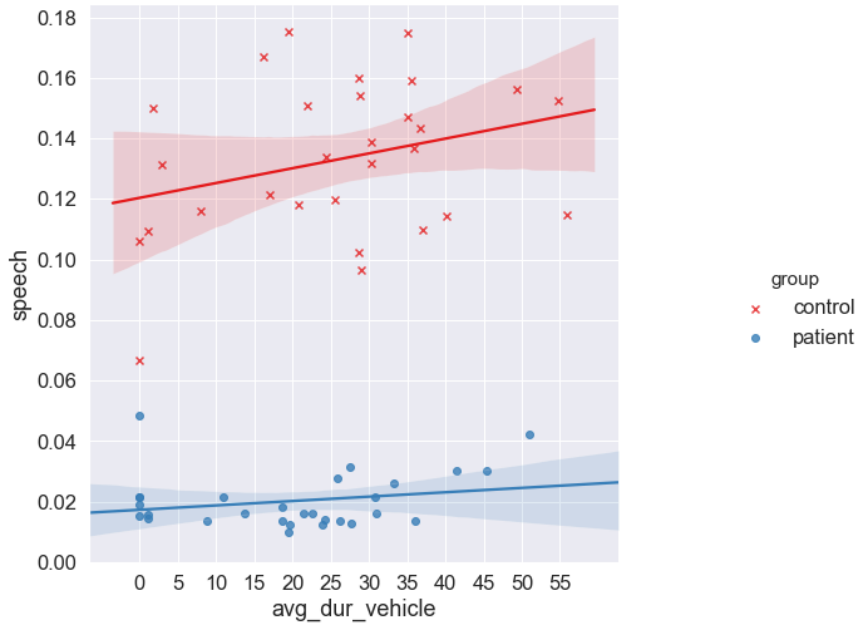

Fig. 5.34 The correlation between average duration of vehicle usage and total percentage of speech produced by the wearer and other speakers for the control subjects (N=29) and the depressed patients (N=29)

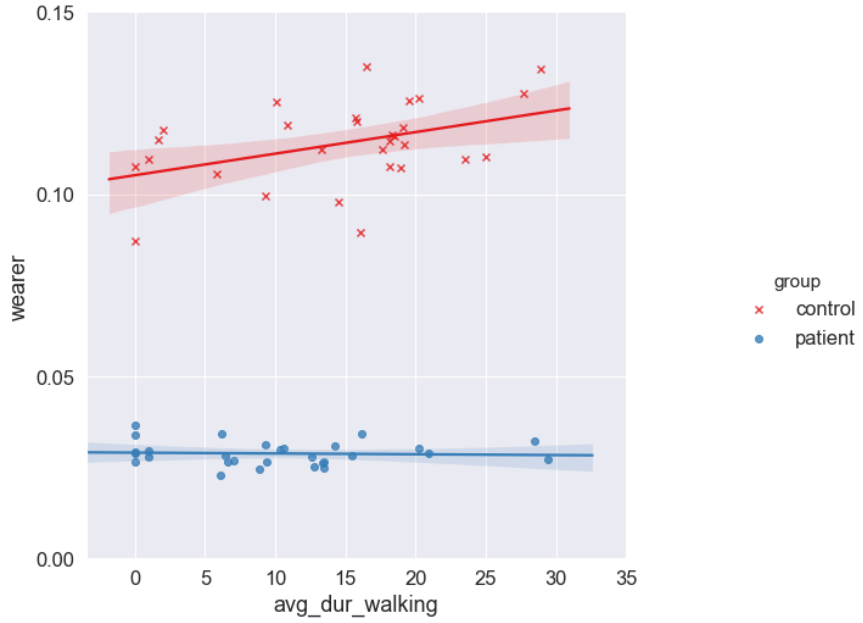

Fig. 5.35 The correlation between average duration of walking and the speech produced by the wearer for the control subjects (N=29) and the depressed patients (N=29)

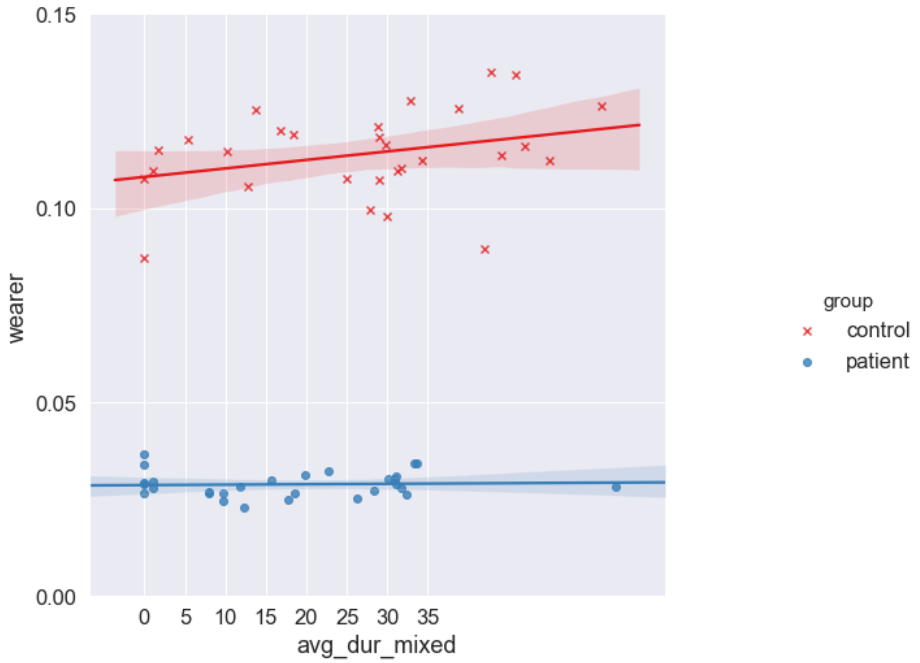

Fig. 5.36 The correlation between average duration of mixed activities and the speech produced by the wearer for the control subjects (N=29) and the depressed patients (N=29)

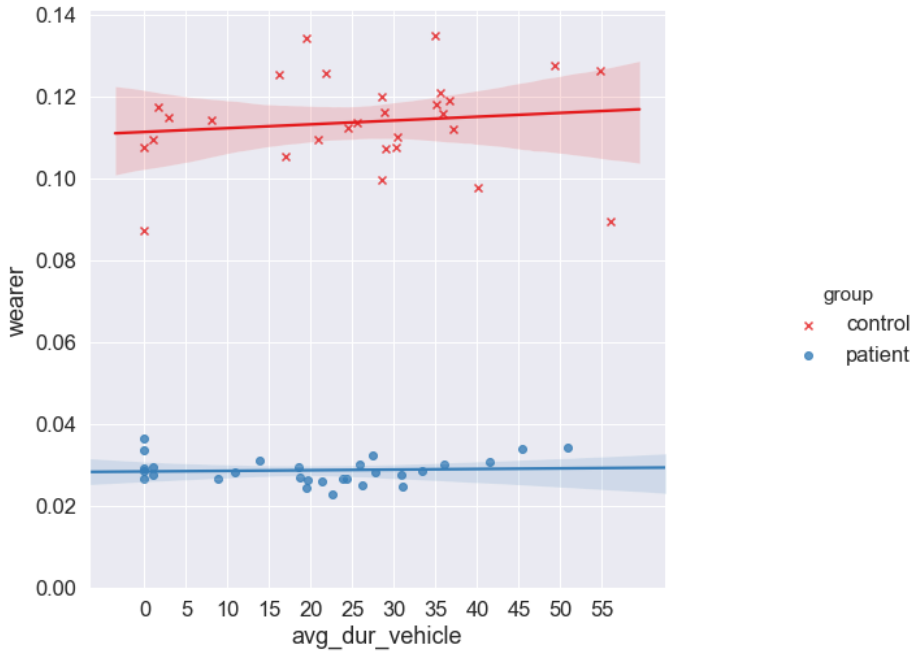

Fig. 5.37 The correlation between average duration of vehicle and the speech produced by the wearer for the control subjects (N=29) and the depressed patients (N=29)

In the following part, I will present new high level features that are extracted from the activity predictions, then show how they can be used as a combination to discriminate between the healthy control group and the LLD patients group. The new features represent more interpretable features than the activity predictions. The features involve:

- the number of bouts for each activity;
- mean of the bout duration for each activity;
- standard deviation of bout duration for each activity;
- percentage of each activity with regards to the total time of the day.

These features are calculated from the class predictions for each activity. Subsequently, these features are fed to the random forest classifier with the two labels (0 for control subject, 1 for depressed patient). I used nested cross validation with 10 stratified folds. The parameter search is performed using the grid search. The parameter ranges are the number of decision trees: [10,250], maximum depth of each tree: [4, 15], maximum number of features: [1, 24]. The ROC curve is shown in Fig. 5.38.

In Table 5.3, the classification report of the new features is shown. The table shows the performance metrics of the random forest model with the best features values: 200 for the number of decision trees, 8 for the depth and 18 for the maximum number of features. The class labels 0 and 1 are for control subjects and LLD patients respectively. As regards the controls, the sensitivity (recall) is 1 which means that the ratio of correctly predicted controls with regards to all the actual controls is perfect. However, the ratio of predicted controls with regards to all the predicted controls is only 0.84. On the other hand, in case of patients, The ratio of correctly predicted patients with regards to all the predicted patients are perfect while there a decline in sensitivity in this case. The average f1-score for the model is 89% and the accuracy is 90% which means about 6 participants are classified incorrectly.

Table 5.3 The classification report for nested cross validation. The input is the new high level features extracted from the activity predictions. The support column shows the number of examples for each class. The macro average represents the unweighted mean of the measure i.e. this does not take the label imbalance into account but the weighted average is the weighted mean taking the class balance into account.

class label	precision	recall	f1-score	support
0	0.84	1.0	0.91	29
1	1.0	0.76	0.86	29
macro avg	0.92	0.88	0.89	58
weighted avg	0.92	0.88	0.89	58

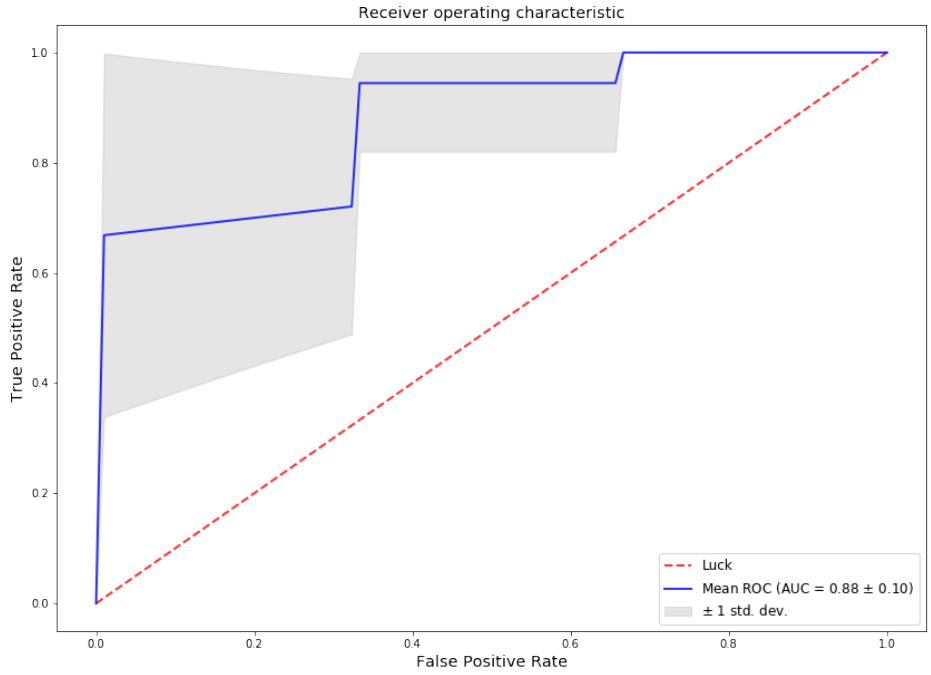

Fig. 5.38 The ROC curve for the new high level features of activities. The blue line represents the mean ROC of the outer k-folds in nested cross validation with 10 different trials. The grey area around the red line represents the standard deviation.

5.4 Conclusion

In this chapter, I present three biomarkers (walking, mixed activity, vehicle usage) from the physical activity that gives a significant separation between the healthy control subjects and LLD patients. The input data in .OMX format is preprocessed and then prepared for the feature extraction and then fed to a pre-trained model developed with Random Forest and Hidden Markov Models. The results show the predictions of each of the activities (sleep, walking, sit.stand, using vehicle, mixed activities and cycling). The sleep patterns, sedentary actions and cycling did not show significance within the groups. On the other hand, walking, mixed activities and vehicle usage show a significance within the groups and then can be used as biomarkers. Nevertheless, these biomarkers do not correlate with any of the key clinical variables used for the Depression diagnosis except for MADRS in LLD group. The activity markers also correlate with the two speech measures for the controls group. In addition, I present a new set of high level features that are more general than specific activity and these features are fed to random forest. The performance shows 89% as F1-score.

Chapter 6

Conclusion

6.1 Summary

Wearable computing technology and machine learning/deep learning are widely used in the social interactions analysis. Unlike the state-of-art social interactions systems especially for healthcare applications which were often applied in lab-settings, this thesis incorporated the social interactions in naturalistic settings. In particular, this thesis strived to verify the research hypothesis that social interactions of LLD could be measured and quantified using wearable sensing platforms and machine learning models. This analysis comprised the verbal interactions analysis and physical activities analysis. The research hypothesis was tested focusing on two main issues: the depression dataset was not annotated for ethical constraints and the data was recorded in naturalistic settings aiming to constructing robust speech models and the discovery of digital markers. This issue prevented using the dataset for training and also made the validation difficult. This was solved by training on another dataset recorded with the same device and the same settings and then evaluating on the depression dataset testing the ability to discriminate between healthy and depressed participants respectively. The second issue of naturalistic settings lead to challenging cases especially in speech where unexpected noise conditions emerged such as audible speech within background and challenging outdoor noises. This was resolved by training a novel deep learning model discussed in Chapter 3.

In response to the research question on detecting speech in naturalistic settings, a robust novel deep learning speech detection model (NatSpeech) was developed. This model predicted all the speech events (segments) for both of the WAM wearers and other people engaging with them. The predicted speech frames were leveraged to compute the average percentage of speech over the week which was exploited as a biomarker and showed a significant difference between controls and cases. However, it did not show any correlation

with the clinical key variables for depression diagnosis. A variety of comparisons and state-of-art datasets were evaluated to show superior performance of the speech detection model discussed in Chapter 3. The output of this model were subsequently used as an input to the second robust novel deep learning model (NatWearer) to predict the speech produced by the wearers of WAM device themselves. Thus, this helped answer and quantify how much and how often the wearer engaged in any conversation. Following that the average predictions of the wearer's speech was also used as a biomarker for depression (Chapter 4). This addressed the second research question and also present another speech biomarker for depression which is the amount of speech produced by the wearer. this indicates that the depressed people were talking less and also engaging with less number of people than the control subjects.

Finally, in Chapter 5, activity markers (average duration of walking, average duration of mixed activities, average duration of vehicle usage) were extracted from the accelerometer data using a pre-trained model and these markers showed association with the depression key variable MADRS. After that new high level features were extracted from the activities predictions and combined to be fed to a random forest model that showed a high performance (0.86% f1-score for predicting the depression and 90% overall accuracy) to discriminate between the control subjects group and depressed patients group. At the end I analysed the association between the activity markers with the two speech markers which showed a significant association between the objective measures.

The implications of this research are presented in two conferences. First, an abstract paper that presents the idea of the application to depression was published in International Meeting of Royal College of Psychiatry, London, UK 2019. Second, the technical machine learning idea was presented as a poster in The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery (ECML-PKDD), Dublin, Ireland 2018. The implications of this research are also that the he speech markers can be replicated in medical research and helped to discriminate between depressed patients and healthy participants. Also the speech measures can be replicated with longitudinal data to predict the onset and progression over time of depression. Furthermore, the speech detection can be replicated to detect the speech and then analyse specific characteristic of the speech such as prosody features that distinguish the depression. Also, the speech detection and wearer detection can be used in speech recognition and speaker recognition systems.

6.2 Limitations

Limitations to this study included cross-sectional design and small sample sizes. The cross-sectional design of the study made it not possible to analyse the implication in the association

between the movement patterns and four key clinical variables of the depression discussed in the thesis. As regards the small sample sizes, using data augmentation to construct a reasonable-size training dataset compared to the depression dataset was not considered. especially the speech data was noisy. Further, adding more clean data and then add the noise artificially was violating the hypothesis of naturalistic settings. So, data augmentation needs to synthesise noisy data from the original dataset without the content i.e. human-like speech but not real. Moreover, in the wearer detection, speaker segmentation was not performed to analyse the amount of speech produced by each speaker in the conversation but only focus on the WAM wearer and did not cluster other speakers and so inferring to how many people the wearer engages with was not possible. As regards the activity predictions, I did not check the feature extraction procedure by making overlap between frames extracted which might affect the performance positively or negatively.

As regards the models trained for speech detection or wearer detection, the models were trained on the discussion dataset and transfer learning concept has been used to apply the pretrained models on the depression dataset. The depression dataset has annotations for whether the participant is a control subject or depressed. However, the models were not trained on the depression dataset directly to create a model that could be optimised to discriminate between the groups.

In the exploratory analysis, multiple comparisons were not adjusted when correlating speech measures with key clinical variables. While the speech classifier and the wearer classifier were accurate in detecting the total speech and wearer respectively, I could not directly generate the accuracy of these classifiers with the depression dataset since listening to and annotating the recordings are not ethically possible. Therefore, how accurate these measures are could not be concluded for this population. Since depression has been associated with abnormalities in specific acoustic features of speech, it was possible that the classifier may have performed differently with the LLD group than controls.

Since groups did not differ in living status, I did not control for this in my analysis. This factor may be particularly important with our measure of speech, since living alone may directly influence the quantity of speech detected. Other factors that we did not control for that may influence the association between social functioning and depression include gender, culture, socio-economic status, and whether participants live in rural, urban or metropolitan areas. Similarly, I did not take into account whether LLD was early-onset or late-onset; these appear to be two distinct types of LLD that have different associations with social functioning. The objective speech measures did not capture qualitative or subjective factors of social interaction, such as satisfaction with social support, which have been shown to be powerful, consistent predictors of depression in older people. Neither do they discriminate

the type of social interactions that are important in LLD, such as emotional and instrumental support.

6.3 Future Work

Future research will include a variant of Wavenet model or other deep learning models suitable for speech synthesis to work for noisy speech conserving the characteristics of the speech. This will create a synthesised data that could be used to augment the original dataset and also could be used for benchmarking and creating pre-trained deep learning models for speech. The idea is that depression dataset is unannotated dataset and WaveNet could be used to synthesise new speech without the content by creating the intonations of the speakers. Therefore, this new synthesised data could be added to the training data and enhance the model performance of the depression dataset. Also, this new data could be used as a benchmark for pre-trained models. Future research will also include analysing the semantics of text from the self-reports by the participants. This also will include the sentiment analysis of the text to produce the sentiment for the participant in each group.

Future research also replicate the findings in this thesis to test external validity and should control for potential confounds such as living status, gender, and culture. It would also be of interest to investigate whether the quantity of speech detected reflects a trait marker of depression or current depressive state. Longitudinal research should measure changes in speech over the onset, course, and remission of depression, and investigate causality and the direction of the relationship between speech and LLD. Methods of detecting more specific variables from this speech data should also be developed, such as measuring acoustic characteristics of the wearer's speech such as prosody.

References

[1] Abowd, D., Orr, A. K., and Brotherton, J. (1998). Context-awareness in wearable and ubiquitous computing. *Virtual Reality*, 3:200–211.

[2] Adaskevicius, R. (2014). Method for recognition of the physical activity of human being using a wearable accelerometer. *Elektronika ir Elektrotechnika*, 20:127–131.

[3] Alickovic, E., Kevric, J., and Subasi, A. (2018). Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. *Biomedical Signal Processing and Control*, 39:94–102.

[4] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Van Essen, B. C., Awwal, A. A., and Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics (Switzerland)*, 8(3):1–67.

[5] Anwary, A., Yu, H., and Vassallo, M. (2018). An automatic gait feature extraction method for identifying gait asymmetry using wearable sensors. *Sensors*, 18(676).

[6] Atallah, L., Lo, B., King, R., and Yang, G. (2011). Sensor positioning for activity recognition using wearable accelerometers. *IEEE Transactions on Biomedical Circuits and Systems*, 5:320–329.

[7] Benjamin, B. (1996). Audio augmented reality: A prototype tour guide. *in proceedings of the ACM conference on Human Factors in Computing Systems*, pages 210–211.

[8] Bennasar, M., Hicks, and Setchi, R. (2015). Feature selection using joint mutual information maximisation. *Expert Systems with Applications*, 42:8520–8532.

[9] Bimbot, F. J., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., and Reynolds, D. A. (2004). A Tutorial on Text-Independent Speaker Verification. *Journal on Applied Signal Processing*, 4:430–451.

[10] Bolón-Canedo, V., Sánchez-Maroño, N., and Alonso-Betanzos, A. (2013). A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34:483–519.

[11] Bonato, P. (2010). Wearable Sensors and Systems From Enabling Technology to Clinical Applications. *IEEE Engineering in Medicine and Biology Magazine*, 29(3):25–36.

[12] Bousquet, L. B. O. (2012). The Tradeoffs of Large Scale Learning". In Sra, Suvrit; Nowozin, Sebastian; Wright, Stephen J. (eds.). Optimization for Machine Learning. *Cambridge: MIT Press.*, pages 351—-368.

[13] Breiman, L. (2001). *Machine Learning*, 45:5–32.

[14] B.S., E. and A., S. (2010). Cambridge Dictionary of Statistics. *Cambridge University Press*.

[15] Bulling, A., Blanke, U., and Schiele, B. (2014). A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys*, 46(3):1–33.

[16] c, C. (2006). *Pattern recognition and machine learning*. Springer. ISBN 0-387-31073-8.

[17] Campbell, J. P. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85(9):1437–1462.

[18] Celis-Morales, C. A., Lyall, D. M., Welsh, P., Anderson, J., Steell, L., Guo, Y., Maldonado, R., Mackay, D. F., Pell, J. P., Sattar, N., and Gill, J. M. R. (2017). Association between active commuting and incident cardiovascular disease, cancer, and mortality: prospective cohort study. *BMJ*, 357.

[19] Chen, G., Wang, A., Zhao, S., Liu, L., and Chang, C. (2017). Latent feature learning for activity recognition using simple sensors in smart homes. *Multimedia Tools and Applications*, pages 1–19.

[20] Chen, K. (2003). Towards better making a decision in speaker verification. *Pattern Recognition*, 36(2):329–346.

[21] Chen, K. and Salman, A. (2011). Learning speaker-specific characteris- tics with a deep neural architecture. *IEEE Transactions on Neural Networks*, 22(11):1744–1756.

[22] Chen, Z., Lin, M., Chen, F., Lane, N. D., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T., and Campbell, A. T. (2013). Unobtrusive sleep monitoring using smartphones. *in the 7th International Conference Pervasive Computing Technologies for Healthcare (PervasiveHealth)*, pages 145–152.

[23] Chengalvarayan, R. (1999). Robust Energy Normalization using Speech/Non-speech Discriminator for German Connected Digit Recognition. *International Speech Communication Association Eurospeech*.

[24] Chernbumroong, S., Cang, S., Atkins, A., and Yu, H. (2013). Elderly activities recognition and classification for applications in assisted living. *Expert Systems with Applications*, 40:1662–1674.

[25] Chernbumroong, S., Cang, S., and Yu, H. (2014). A practical multi-sensor activity recognition system for home-based care. *Decision Support Systems*, 66:61–70.

[26] Chernbumroong, S., Cang, S., and Yu, H. (2015). Genetic algorithm-based classifiers fusion for multisensor activity recognition of elderly people. *IEEE Journal of Biomedical and Health Informatics*, 19:282–289.

[27] Choudhury, T. and Pentland, A. (2003). Sensing and modeling human networks using the sociometer. pages 216–222.

[28] Ciresan, D., Meier, U., and Schmidhuber, J. (June 2012). Multi-column deep neural networks for image classification. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3642–3649.

[29] Colbert, L., Matthews, C., Havighurst, T., Kim, K., and Schoeller, D. (2011). Comparative validity of physical activity measures in older adults. *Medicine Science in Sports and Exercise*, 43:867–876.

[30] Cook, A., Gargiulo, G., and Hamilton, T. L. (2015). Open platform, eight-channel, portable bio-potential and activity data logger for wearable medical device development. *Electronics Letters*, 51:1641–1643.

[31] Cortes, C. and Vapnik, V. N. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

[32] Dahl, G. E., Yu, D., Member, S., Deng, L., and Acero, A. (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 20(1):30–42.

[Dauphin et al.] Dauphin, Y. N., de Vries, H., Chung, J., and Bengio, Y. Rmsprop and equilibrated adaptive learning rates for non-convex optimization.

[34] Dauphin, Y. N., Fan, A., Auli, M., and Grangier, D. (2017). *Language Modeling with Gated Convolutional Networks*. ICML'17. JMLR.org.

[35] Davis, K., Owusu, E., Bastani, V., Marcenaro, L., Hu, J., Regazzoni, C., and Feijs, L. (2016). Activity recognition based on inertial sensors for ambient assisted living. *IEEE international conference on formation fusion (FUSION)*, pages 371–378.

[36] Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7:1–30.

[37] Doherty, A., Jackson, D., Hammerla, N., Plötz, T., Olivier, P., Granat, M. H., White, T., van Hees, V. T., Trenell, M. I., Owen, C. G., Preece, S. J., Gillions, R., Sheard, S., Peakman, T., Brage, S., and Wareham, N. J. (2017). Large Scale Population Assessment of Physical Activity Using Wrist Worn Accelerometers: The UK Biobank Study. *PLOS ONE*, 12(2):e0169649.

[38] Doherty, A., Smith-Bryne, K., Ferreira, T., Holmes, M., Holmes, C., Pulit, S., and Lindgren, C. (2018). Gwas identifies 14 loci for device-measured physical activity and sleep duration. *Nature Communications*, 9.

[39] Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communications*, 40(1):33–60.

[40] Du, B., Xiong, W., Wu, J., Zhang, L., Zhang, L., and Tao, D. (2017). Stacked convolutional denoising auto-encoders for feature representation. *IEEE Transactions on Cybernetics*, 47(4):1017–1027.

[41] Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

[42] Dumoulin, V. and Visin, F. (2016). A guide to convolution arithmetic for deep learning.

[43] Eclipse Deeplearning4j Development Team (2016). DL4J: Deep Learning for Java.

[44] Eyben, F., Weninger, F., Squartini, S., and Schuller, B. (2013). Real-life voice activity detection with LSTM Recurrent Neural Networks and an application to Hollywood movies. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 1(3):483–487.

[45] Filippoupolitis, A., Oliff, W., Takand, B., and Loukas, G. (2017). Location-enhanced activity recognition in indoor environments using off the shelf smart watch technology and ble beacons. *Sensors*, 17(1230).

[46] Fiske, A., Wetherell, J., and Gatz, M. (2009). Depression in older adults. *Annual Review of Clinical Psychology*, 5:63—-89.

[47] Fukuda, T., Ichikawa, O., and Nishimura, M. (2010). Long-term spectro-temporal and static harmonic features for voice activity detection. *IEEE Journal of Selected Topics in Signal Processing*, 4:834–844.

[48] Fukushima, K. N. (1988). A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1:119–130.

[49] Gallego, J. A., Rocon, E., Koutsou, A. D., and Pons, J. L. (2011). Analysis of kinematic data in pathological tremor with the Hilbert-Huang transform. *5th International IEEE/EMBS Conference on Neural Engineering, NER 2011.*, pages 80–83.

[50] Ganapathy, S., Rajan, P., and Hermansky, H. (2011). Multi-layer Perceptron based Speech Activity Detection for Speaker Verification. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*.

[51] Ganchev, T. N., Fakotakis, N., and Kokkinakis, G. (2005). Comparative Evaluation of Various MFCC Implementations on the Speaker Verification Task. *Proceedings of 10th International Conference on Speech and Computer*, 2:191–194.

[52] Gelly, G. and Gauvain, J.-L. (2018). Optimisation of RNN-Based Speech Activity Detection. *IEEE/ACM Transactions on Audio Speech and Language Processing*, 26(3):646–656.

[53] Goldberg, Y. (2017). *Neural Network Methods in Natural Language Processing*. Morgan and Claypool Publishers.

[54] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

[55] Goodfellow, I., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013). Maxout networks. volume 28 of *Proceedings of Machine Learning Research*, pages 1319–1327, Atlanta, Georgia, USA. PMLR.

[56] Google (25 May 2017). Google's alphago ai wins three-match series against the world's best go player.

[57] Gorriz, J., Ramirez, J., Lang, E., and Puntonet, C. (2005). Hard c-means clustering for voice activity detection. *Speech Communications*, 48(12):1638–1649.

[58] Graves, A. (2012). Supervised Sequence Labelling with Recurrent Neural Networks. *Textbook, Studies in Computational Intelligence, Springer*.

[59] Graves, A. and Schmidhuber, J. (July-August 2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks, Elsevier*, 18(5-6):602–610.

[60] Guo, J., Zhou, X., Sun, Y., Ping, G., Zhao, G., and Li, Z. (2016). Smartphone-based patients' activity recognition by using a self-learning scheme for medical monitoring. *Journal of Medical Systems*, 40(140).

[61] Hammerla, N., Halloran, S., and Ploetz, T. (2016). Deep convolutional and recurrent models for human activity recognition using wearables.

[62] Hassan, M., Uddin, M. Z., Mohamed, A., and Almogren, A. (2018). A robust human activity recognition system using smartphone sensors and deep learning. *Future Generation Computer Systems*, 81:307–313.

[63] He, Z. and Jin, L. (2009). Activity recognition from acceleration data based on discrete consine transform and SVM. *IEEE international conference on Systems man and cybernetics*, pages 5041–5044.

[64] Hees, V., Renstrom, F., Wright, A., Gradmark, A., Catt, M., Chen, K. Y., Löf, M., Bluck, L., Pomeroy, J., Wareham, N. J., Ekelund, U., Brage, S., and Franks, P. W. (2011). Estimation of daily energy expenditure in pregnant and non-pregnant women using a wrist-worn tri-axial accelerometer. *PLoS ONE*, 6.

[65] Hermasnsky, H. (1990). Perceptual Linear Predictive Analysis of Speech. *Journal of the Acoustical Society of America*, pages 1738–1752.

[66] Hinton, G. (2012). Neural networks for machine learning.

[67] Hinton, G. and McClelland, J. (1988). Learning representations by recirculation. *In Proceeding of Annual Conference on Neural Information Processing Systems NIPS'1987*, pages 358–366.

[68] Hinton, G. and Sejnowski, T. (1999). Unsupervised learning: Foundations of neural computation. *MIT Press*.

[69] Hinton, G. and Zemel, R. (1994). Autoencoders, minimum description length, and Helmholtz free energy. *In Proceeding of Annual Conference on Neural Information Processing Systems(NIPS'1993)*.

[70] Hinton, G. E. and Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507.

[71] Hirsch, H. G. (2005). FaNT-Filtering and Noise Adding Tool.

[72] Hochreiter, S., Bengio, Y., Frasconi, P., and Schmidhuber, J. (2001). *Gradient flow in recurrent nets: the difficulty of learning long-term dependencies*.

[73] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

[74] Hodgetts, S., Gallagher, P., Stow, D., Ferrier, N. I., and O'Brien, J. T. (2017). The impact and measurement of social dysfunction in late-life depression: An evaluation of current methods with a focus on wearable technology. *International Journal of Geriatric Psychiatry*, 32:247–255.

[75] Hu, L., Chen, Y., Wang, S., and Chen, Z. (2014). b-COELM: A fast, lightweight and accurate activity recognition model for mini-wearable devices. *Pervasive and Mobile Computing*, 15:200–214.

[76] Hughes, T. and Mierle, K. (2013). Recurrent neural networks for voice activity detection. *Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing (ICASSP)*, pages 7378–7382.

[77] Hyvärinen, A. (2005). Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709.

[78] Ignatov, A. (2018). Real-time human activity recognition from accelerometer data using Convolutional Neural Networks. *Applied Soft Computing*, 62:915–922.

[79] Kalantarian, H., Alshurafa, N., Le, T., and Sarrafzadeh, M. (2015). Monitoring eating habits using a piezoelectric sensor-based necklace. *Computers in Biology and Medicine*, 58:46–55.

[80] Kanai, Y., Morita, S., and Unoki, M. (2013). Concurrent processing of voice activity detection and noise reduction using empiri- cal mode decomposition and modulation spectrum analysis. *in proceedings of INTERSPEECH*, pages 742–746.

[81] Kawamoto, K., Tanaka, T., and Kuriyama, H. (2014). Your activity tracker knows when you quit smoking. *in Proceedings of the 2014 ACM International Symposium on Wearable Computers - ISWC '14*, pages 107–110.

[82] KEB, O., Lech, M., and Allen, N. (2014). Prediction of major depression in adolescents using an optimized multi-channel weighted speech classification system. *Biomedical Signal Processing Control*, 14:228–239.

[83] Kemp, J., Gaura, E., Rednic, R., and Brusey, J. (2013). Long-term behavioural change detection through pervasive sensing. *in Proceeding of 14th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD 2013), Honolulu, Hawaii, U.S.A*, pages 629–634.

[84] Kim, H., Jin, K. H., Jihyun, L., Sanghyun, Y., Kyoung-Gu, W., Sung, N. J., and Seungmin, Y. (2013). Monitoring for disease progression via mathematical time-series modeling: actigraphy-based monitoring patients with depressive disorder. *In IEEE 10th Consumer Communications and Networking Conference (CCNC), Las Vegas*, pages 56–61.

[85] Kim, J., Kim, J., Lee, S., Park, J., and Hahn, M. (2016a). Vowel based Voice Activity Detection with LSTM Recurrent Neural Network. *Proceedings of the 8th International Conference on Signal Processing Systems (ICSPS 2016)*, pages 134–137.

[86] Kim, S. K., Park, Y. J., and Lee, S. (2016b). Voice activity detection based on deep belief networks using likelihood ratio. *Journal of Central South University*, 23(1):145–149.

[87] Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

[88] Kitaoka, N., Yamada, T., Tsuge, S., Miyajima, C., Yamamoto, K., Nishiura, T., Nakayama, M., Denda, Y., Fujimoto, M., Takiguchi, T., Tamura, S., Matsuda, S., Ogawa, T., Kuroiwa, S., Takeda, K., and Nakamura, S. (2009). CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments. *Acoustical Science and Technology*, 30(5):363–371.

[89] Kotti, M., Moschou, V., and Kotropoulos, C. (2008). Speaker segmentation and clustering. *Signal processing*, 88(5):1091–1124.

[90] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). Imagenet classification with deep convolutional neural networks. *NIPS 2012: International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada*.

[91] Kwon, Y., Kang, and Bae, C. (2014). Unsupervised learning for human activity recognition using smartphone sensors. *Expert Systems with Applications*, 41:6067–6074.

[92] K.Woo, Yang, T., Park, K., and Lee, C. (2000). A Robust Voice Activity Detection Algorithm for Estimating Noise Spectrum. *IEEE Electronics Letters*.

[93] Laerhoven, K. V., Borazio, M., Kilian, D., and Schiele, B. (2008). Sustained logging and discrimination of sleep postures with low-level, wrist- worn sensors. *In 12th IEEE International Symposium on Wearable Computers*, pages 69–76.

[94] Laudanski, A., Brouwer, B., and Li, Q. (2015). Activity classification in persons with stroke based on frequency features. *Medical Engineering & Physics*, 37.

[95] LeCun, Y. (1987). *Modèles connexionistes de l'apprentissage*. Phd thesis, Université de Paris.

[96] LeCun, Y. (1989). Generalization and network design strategies. *Technical ReportCRG-TR-89-4, University of Toronto*.

[97] Leglaive, S., Hennequin, R., and Badeau, R. (2015). Singing voice detection with deep recurrent neural networks. *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 121–125.

[98] Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R., and Tang, J. (2017). Feature selection: A data perspective. *ACM Computing Surveys (CSUR)*, 50.

[99] Liang, H., Hu, B., Liu, Z., and Yan, L. (2017). Investigation of different speech types and emotions for detecting depression using different classifiers. *Speech Communications*, 90:39—46.

[100] Lin, Y.-P. and Jung, T.-P. (2017). Improving EEG-Based Emotion Classification Using Conditional Transfer Learning. *Frontiers in Human Neuroscience*, 11(334).

[101] Liu, L., Peng, Y., Liu, M., and Huang, Z. (2015). Sensor-based human activity recognition system with a multilayered model using time series shapelets. *Knowledge-Based Systems*, 90:138–152.

[102] Liu, S., Gao, R., John, D., and Freedson, J. S. P. (2012). Multi sensor data fusion for physical activity assessment. *IEEE Transactions on Biomedical Engineering*, 59:687–696.

[103] Lu, X., Unoki, M., Isotani, R., Kawai, H., and Nakamura, S. (2011). Adaptive regularization framework for robust voice activity detection. *in proceedings of INTERSPEECH*, pages 2653–2653.

[104] Machado, I., Gomes, A., Gamboa, H., Paixão, V., and Costa, R. (2015). Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization. *Information Processing and Management*, 51:204–214.

[105] Mannini, A., Stephen, M. R., Intille, S., Sabatini, A. A. M., and Haskell, W. (2014). Activity recognition using a single accelerometer placed at the Wrist or Ankle. *Medicine & Science in Sports & Exercise*, 45(11):2193–2203.

[106] Masci, J., Meier, U., Cireşan, D., and Schmidhuber, J. (2011). *Stacked Convolutional Auto-Encoders for Hierarchical Feature Extraction*. Springer Berlin Heidelberg, Berlin, Heidelberg.

[107] Matthews, M., Abdullah, S., Gay, G., and Choudhury, T. (2014). Tracking mental well-being: Balancing rich sensing and patient needs. *IEEE Computer Society*, 47(4):36–43.

[108] McCulloch, W. S. and Pitts, W. (1988). A logical calculus of the ideas immanent in nervous activity. pages 15–27.

[109] Mehrang, S., Pietila, J., Tolonen, J., Helander, E., Jimison, H., Pavel, M., and Korhonen, I. (2017). Human Activity Recognition Using A Single Optical Heart Rate Monitoring Wristband Equipped with Triaxial Accelerometer. *in proceedings of joint conference of the European Medical and Biological Engineering Conference (EMBEC) and the Nordic-Baltic Conference on Biomedical Engineering and Medical Physics (NBC)*, pages 587–590.

[110] Mendelev, V. S., Prisyach, T. N., and Prudnikov, A. A. (2015). Robust Voice Activity Detection with Deep Maxout Neural Networks. *Modern Applied Science*, 9(8):153–159.

[111] Min, J.-K., Doryab, A., Wiese, J., Amini, S., Zimmerman, J., and Hong, J. I. (2014). Toss 'n' turn: smartphone as sleep and sleep quality detector. *In proceedings of the 32nd annual ACM conference on Human factors in computing systems*, pages 477–486.

[112] Moncada-Torres, A., Leuenberger, K., Gonzenbach, R., Luft, A., and Gassert, R. (2014). Activity classification based on inertial and barometric pressure sensors at different anatomical locations. *Physiological Measurement*, 35(1245).

[113] Montalto, F., Guerra, C., Bianchi, V., Munari, I. D., and Ciampolini, P. (2015). MuSA: Wearable Multi Sensor Assistant for Human Activity Recognition and Indoor Localization. *In Ambient Assisted Living, Springer*, pages 81–92.

[114] Mortazavi, B., Pourhomayoun, M., Alsheikh, G., Alshurafa, N., Lee, S., and Sarrafzadeh, M. (2014). Determining the single best axis for exercise repeti- tion recognition and counting on smartwatches. *In Proceedings of 11th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 33–38.

[115] Mundt, J. C., Vogel, A. P., Feltner, D., and Lenderking, W. R. (2012). Vocal acoustic biomarkers of depression severity and treatment response. *Biological Psychiatry*, 72:580–587.

[116] Nair, V. and Hinton, G. (June 2010). Rectified linear units improve restricted boltzmann machines. *In Proceedings of the 27th International Conference on Machine Learning (ICML-10), Haifa, Israel*, pages 807–814.

[117] Nemer, E., Goubran, R., and Mahmoud, S. (2000). Robust Voice Ac- tivityDetection usingHigher-order Statistics in the LPCResid- ual Domain. *IEEE Electronics Letters*.

[118] Nesterov, Y. (2004). Introductory lectures on convex optimization : a basic course. Applied optimization.

[119] Nguyen, K. D., Chen, I. M., Luo, Z., Yeo, S. H., and Duh, H. B. L. (2011). A wearable sensing system for tracking and monitoring of functional arm movement. *IEEE/ASME Transactions on Mechatronics*, 16(2):213–220.

[120] Nwe, T. L., Sun, H., Ma, B., Member, S., Li, H., and Member, S. (2012). Speaker Clustering and Cluster Purification Methods for RT07 and RT09 Evaluation Meeting Data. *IEEE Transactions on Audio Speech and Language Processing*, 20(2):461–473.

[121] O'Brien, J. T., Gallagher, P., Stow, D., Hammerla, N., Ploetz, T., Firbank, M., Ladha, C., Ladha, K., Jackson, D., McNaney, R., Ferrier, I. N., and Olivier, P. (2017). A study of wrist-worn activity measurement as a potential real-world biomarker for late-life depression. *Psychological Medicine*, 47(1):93–102.

[122] Obuchi, Y. (2016). Framewise speech-nonspeech classification by neural net- works for voice activity detection with statistical noise suppression. *Proceedings of IEEE International Conference on Acoustics and Speech Signal Processing (ICASSP)*, pages 5715–5719.

[123] Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio.

[124] Pang, I., Okubo, Y., Sturnieks, D., Lord, S., and Brodie, A. (2019). Detection of near falls using wearable devices: a systematic review. *Journal of geriatric physical therapy*, 42:48–56.

[125] Pantelopoulos, A. and Bourbakis, N. G. (2010). A Survey onWearable Sensor-Based Systems for Health Monitoring and Prognosis. *IEEE Transactions on System, Man and Cybernetics Applications and Reviews*, 40(1):1–12.

[126] Panwar, M., Dyuthi, S., Prakash, K., Biswas, D., Acharyya, A., Maharatna, K., Gautam, A., and Naik, G. (2017). CNN based approach for activity recognition using a wrist-worn accelerometer. *IEEE Annual International Conference in Engineering in Medicine and Biology Society (EMBC)*, pages 2438–2441.

[127] Pavey, T., Gilson, N., Gomersall, S., Clark, B., and Trost, S. (2017). Field evaluation of a random forest activity classifier for wrist-worn accelerometer data. *Journal of Science and Medicine in Sport*, 20:75–80.

[128] Pearce, D. and Hirsch, H. (2000). The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions. *in Proceedings of the Sixth International Conference on Spoken Language Processing, ICSLP 2000*, pages 29–32.

[129] Pearce, D. and Picone, J. (2002). Aurora Working Group: DSR font end LVCSR Evaluation AU/384/02. *Institute for Signal and Information Processing*.

[130] Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1—-17.

[131] Prince, S., Adamo, K., Hamel, M., Hardt, J., Gorber, S., and Tremblay, M. (2008). A comparison of direct versus self-report measures for assessing physical activity in adults: A systematic review. *International Journal of Behavioural Nutrition and Physical Activity*, 5:1—24.

[132] Rabiner, L. and Juang, B. (1986). An introduction to hidden Markov models. *IEEE Acoustics, Speech and Signal Processing Newsletter (IEEE ISSAP)*, 3:4–16.

[133] Ramirez, J., Segura, J. C., Bentitez, C., Garcia, L., and Rubio, A. (2005). Statistical Voice Activity Detection using Multiple Observation Likelihood Ratio Test. *IEEE Signal Processing Letters*, 12(10):689–692.

[134] Ranzato, M., Y., Y. B., and LeCun (2008). Sparse feature learning for deep belief networks. *In Proceeding of Annual Conference on Neural Information Processing Systems NIPS'2007*, pages 1185–1192.

[135] Reyes-Ortiz, J., Oneto, L., Sama, A., Parra, X., and Anguita, D. (2016). Transition-aware human activity recognition using smartphones. *Neurocomputing*, 171:754–767.

[136] Ronao, C. and Cho, S. (2016). Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Systems with Applications*, 59:235–244.

[137] Rouvier, M., Bousquet, P. M., and Favre, B. (2015). Speaker diarization through speaker embeddings. *in Proceedings of 23rd IEEE European Signal Processing Conference (EUSIPCO)*, pages 2082–2086.

[138] Sabia, S., van Hees, V., Shipley, M., Trenell, M., Hagger-Johnson, G., Elbaz, A., Kivimaki, M., and Singh-Manoux, A. (2014). Association between questionnaire and accelerometer-assessed physical activity: the role of sociodemographic factors. *American Journal of Epidemiology*, 179:781–790.

[139] Sadjadi, S. O. and H.L.Hansen, J. (2013). Unsupervised speech activity detection using voicingmeasures and perceptual spectral flux. *IEEE Signal Processing Letters*, 20(3):197–200.

[140] Salcedo-Bernal, A., Villamil-Giraldo, M., and Moreno-Barbosa, A. (2016). Clinical data analysis: An opportunity to compare machine learning methods. *Procedia Computer Science*, 100:731 – 738. International Conference on ENTERprise Information Systems/International Conference on Project MANagement/International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2016.

[141] Sangwan, A., Zhu, W. P., and Ahmad, M. O. (2007). Design and performance analysis of Bayesian, NeymanPearson, and competitive NeymanPearson voice activity detectors. *IEEE Transactions on Signal Processing*, 55(9):4341–4353.

[142] Santini, Z., Koyanagi, A., Tyrovolas, S., and Mason, C. (2015). The association between social relationships and depression: A systematic review. *Journal of Affective Disorders*, 175:53–65.

[143] Scholkopf, B. and Smola, A. J. (2001). Learning with kernels: Support vector machines, regularization, optimization, and beyond. *MIT Press, Cambridge, MA, USA*.

[144] Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45:2673–2681.

[145] Scibelli, F., Roffo, G., Tayarani, M., Bartoli, L., Mattia, G. D., Esposito, A., and Vinciarelli, A. (2018). Depression Speaks: Automatic discrimination between depressed and non-depressed speakers based on non-verbal speech features. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6842—6846.

[146] Sehgal, A. and Kehtarnavaz, N. (2018). A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection. *IEEE Access*, 6:9017–9026.

[147] Seiter, J., Derungs, A., Schuster-Amft, C., Amft, O., and Tröster, G. (2014). Activity Routine Discovery in Stroke Rehabilitation Patients without Data Annotation. *in proceedings of the 8th International Conference on Pervasive Computing Technologies for Healthcare*, 5:270–273.

[148] S.Garofolo, J., Lamel, L., Fisher, W. M., and Fiscus, J. G. (1993). Timit acoustic-phonetic continuous speech corpus ldc93s1. *Philadelphia: Linguistic Data Consortium*.

[149] Shan, Z., Ma, H., Xie, M., Yan, P., Guo, Y., Bao, W., Rong, Y., Jackson, C., Hu, F., and Liu, L. (2015). Sleep Duration and Risk of Type 2 Diabetes: A Meta-analysis of Prospective Studies. *Diabetes Care*, 38.

[150] Sincy, T. V., Sreekumar, K. T., Santhosh, K. C., and Reghu, R. P. C. (2015). Random forest algorithm for improving the performance of speech/non-speech detection. *First International Conference on Computational Systems and Communications (ICCSC)*.

[151] Smith, L. N. (2015). Cyclical learning rates for training neural networks.

[152] Sohn, J., Kim, N. S., and Sung, W. (1999). A statitical model-based voice activity detection. *IEEE Signal Processing Letters*, 6(1):1–3.

[153] Sokolova, M., Japkowicz, N., and Szpakowicz, S. (2006). Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. *Astralasian joint conference on Artificial Intelligence, Springer*, pages 1051–1021.

[154] Sprint, G., Weeks, D., Borisov, V., and Cook, D. (2014). Wearable sensors in ecological rehabilitation environments. *in proceedings of the ACM Intentional Joint Conference on Pervasive and Ubiquitous Computing (UbiComp)*, pages 163–166.

[155] Steenken, A. (1993). Assessment for Automatic Speech Recognition: II. NOISEX-92: A database and An Experiment to Study the effect of additive noise on Speech Recognition Systems. *Speech Communication*, 12(3):247–251.

[156] Steger, M. and Kashdan, T. B. (2009). Depression and Everyday Social Activity, Belonging, and Well-Being. *Journal of counseling psychology*, 56(2):289–300.

[157] Suh, Y. and Kim, H. (2012). Multiple Acoustic Model-Based Discriminative Likelihood Ratio Weighting for Voice Activity Detection. *IEEE Signal Processing Letters*, 19(8):507–510.

[158] Sukittanon, S., Surendran, A. C., Platt, J., Burges, C. J., and Look, B. (2004). Convolutional Networks for Speech Detection. *in proceedings of INTERSPEECH*, pages 2–5.

[159] Suto, J., Oniga, S., Lung, C., and Orha, I. (2017). Recognition rate difference between real-time and offline human activity recognition. *in IEEE International Conference on Internet of Things for the Global Community (IoTGC) 2017*, pages 1–6.

[160] Suto, J., Oniga, S., and Sitar, P. (2016). Feature analysis to human activity recognition. *International Journal of Computers Communications and Control*, 12:116–130.

[161] Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. *ICML'13: Proceedings of the 30th International Conference on International Conference on Machine Learning*, 28:1139–1347.

[162] Tackman, A., Sbarra, D., Carey, A., Donnellan, M., Horn, A., Holtzman, N., Edwards, T., Pennebaker, J., and Mehl, M. (2019). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of personality and social psychology*, 116:817–834.

[163] Teng, P. and Jia, Y. (2013). Voice activity detection via noise reducing using non-negative sparse coding. *IEEE Signal Processing Letters*, 20(5):475–478.

[164] Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*.

[165] Thomas, A., Gallagher, P., Robinson, L., Porter, R., Young, A., Ferrier, N., and O'Brien, J. (2009). A comparison of neurocognitive impairment in younger and older adults with major depression.

[166] Thomas, S., Ganapathy, S., Saon, G., and Soltau, H. (2014). Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions. *Proceeding of IEEE International Conference in Acoustics and Speech Signal Processing (ICASSP)*, pages 2519–2523.

[167] Uddin, M., Salem, A., Nam, I., and Nadeem, T. (2015). Wearable sensing framework for human activity monitoring. *In Proceedings of the ACM workshop on Wearable Systems and Applications*, pages 21–26.

[168] Uray, M., Skocaj, D., Roth, P., Bischof, H., and Leonardis, A. (2007). Incremental LDA learning by combining reconstructive and discriminative approaches. *In BMVC*, pages 272–281.

[169] Uslu, G., Dursunoglu, H., Altun, O., and Baydere, S. (2013). Human activity monitoring with wearable sensors and hybrid classifiers. *International Journal of Computer Information Systems and Industrial Management Applications*, 5:345–353.

[170] Varela, O., San-Segundo, R., and Hern, L. (2011). Combining pulse-based features for rejecting far-field speech in a HMM-based voice activity detector. *Computers and Electrical Engineering*, 37:589–600.

[171] Vinciarelli, A., Pantic, M., and Bourlard, H. (2009). Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759.

[172] Voleti, R., Woolridge, S., Liss, J., and Milanovic, M. (2019). Objective Assessment of Social Skills Using Automated Language Analysis for Identification of Schizophrenia and Bipolar Disorder.

[173] Wang, J., Chen, Y., Hao, S., Peng, X., and Hu, L. (2018). Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters*.

[174] Wang, R., Chen, F., Chen, Z., and Li, T. (2014). Studentlife: assessing mental health, academic performance and behavioural trends of college students using smartphones. *in Proceedings of the ACM Conference on Ubiquitous Computing*.

[175] Wang, Y., Cang, S., and Yu, H. (2019). A survey on wearable sensor modality centred human activity recognition in health care. *Expert Systems with Applications*, 137:167–190.

[176] Wilde, N., Hänsel, K., Haddadi, H., and Alomainy, A. (2015). Wearable Computing for Health and Fitness: Exploring the Relationship between Data and Human Behaviour. pages 1–23.

[177] Willetts, M., Hollowell, S., Aslett, L., Holmes, C., and Doherty, A. (2018). Statistical machine learning of sleep and physical activity phenotypes from sensor data in 96,220 UK Biobank participants. *Scientific Reports Nature*, 8(1):1–10.

[178] William, H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. (2007). Numerical recipes: The art of scientific computing (3rd ed.). *Cambridge University Press*.

[179] Williams, G. (2011). Data mining with Rattle and R: the art of excavating data for knowledge discovery. *Springer Science and Business Media*, 29.

[180] Wu, J., Wang, J., and Liu, L. (2007). Feature extraction via KPCA for classification of gait patterns. *Human Movement Science*, 26:393–411.

[181] Wu, J., Zhang, X., and Member, S. (2011). Efficient Multiple Kernel Support Vector Machine Based Voice Activity Detection. *IEEE Signal Processing Letters*, 18(8):466–469.

[182] Wu, J. and Zhang, X. L. (2011). Maximum Margin Clustering Based Statistical VAD With Multiple Observation Compound Feature. *IEEE Signal Processing Letters*, 18(5):283–286.

[183] Y, Y. Y., Fairbairn, C., and Cohn, J. (2013). Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 4:142–150.

[184] Yella, S. H. and Stolcke, A. (2015). A comparison of neural network feature transforms for speaker diarization. *INTERSPEECH*, pages 3026–3030.

[185] Ying, D., Yan, Y., Dang, J., and Soong, F. (2011). Voice Activity Detection based on unsupervised learning framework. *IEEE transactions on Audio, Speech and Language Processing*, 19(8):2624–2644.

[186] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? *CoRR*, abs/1411.1792.

[187] Zhang, X.-L. and Wang, D. (2016). Boosting Contextual Information for Deep Neural Network Based Voice Activity Detection. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24(2):252–264.

[188] Zhang, X.-l. and Wu, J. (2013a). Deep Belief Networks Based Voice Activity Detection. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(4):697–710.

[189] Zhang, X. L. and Wu, J. (2013b). Denoising Deep Neural Networks Based Voice Activity Detection. *IEEE International Conference of Acoustics, Speech and Signal Processing*.

[190] Zolfaghari, S. and Keyvanpour, M. (2016). SARF: Smart activity recognition framework in ambient assisted living. *In IEEE Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 1435–1443.

[191] Özkanca, Y., Demiroglu, C., Besirli, A., and Selime, C. (2018). Multi-Lingual Depression-Level Assessment from Conversational Speech Using Acoustic and Text Features. *in Proceeding INTERSPEECH 2018, ISCA*, page 3398–3402.

Appendix A

Thesis Appendix

A.1 ROC-curves for speech detection on discussion dataset

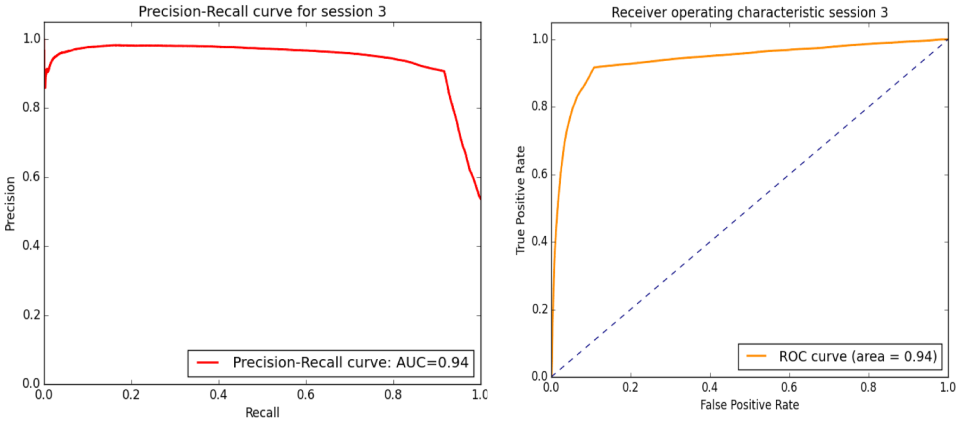

Fig. A.1 ROC and PR curves for session 3 (Discussion Dataset)

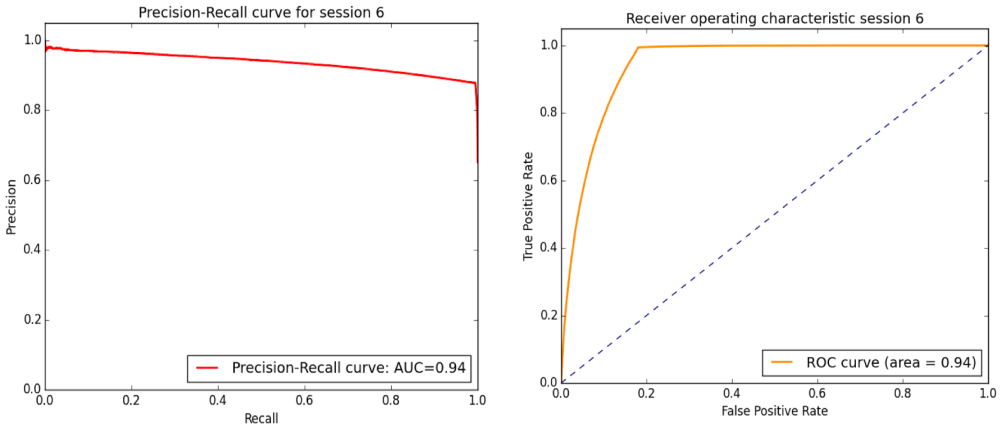

Fig. A.2 ROC and PR curves for session 6 (Discussion Dataset)

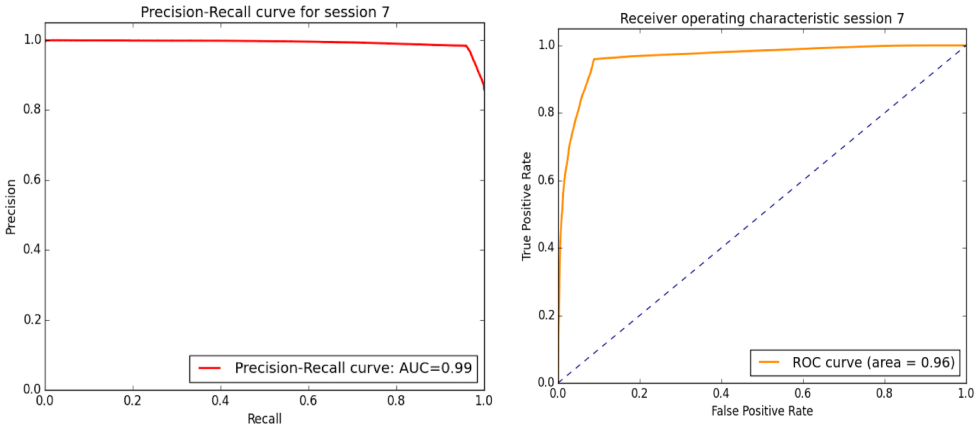

Fig. A.3 ROC and PR curves for session 7 (Discussion Dataset)

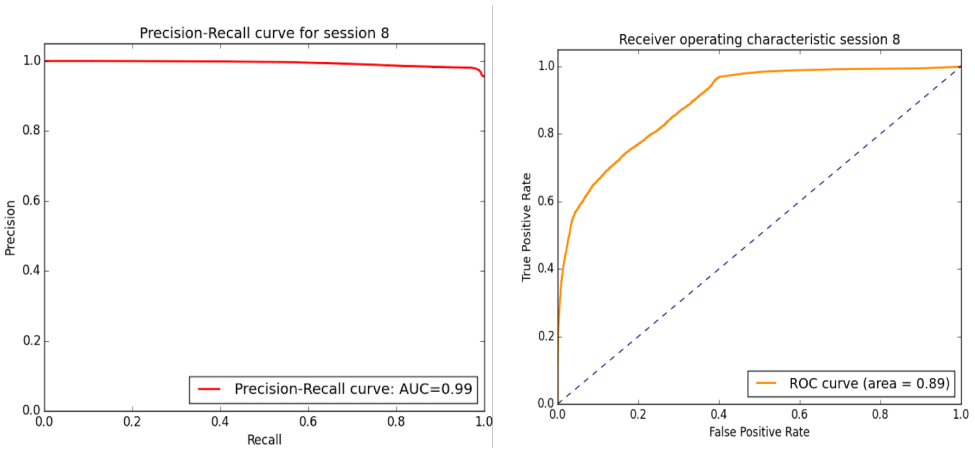

Fig. A.4 ROC and PR curves for session 8 (Discussion Dataset)

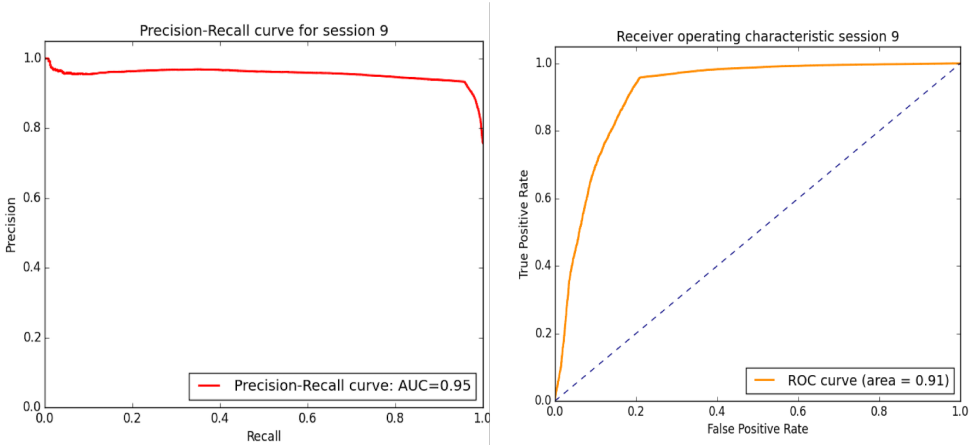

Fig. A.5 ROC and PR curves for session 9 (Discussion Dataset)

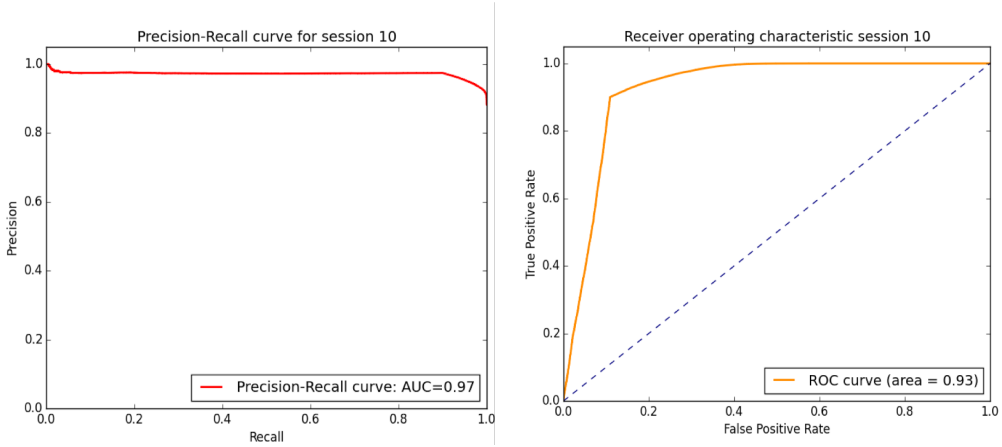

Fig. A.6 ROC and PR curves for session 10 (Discussion Dataset)

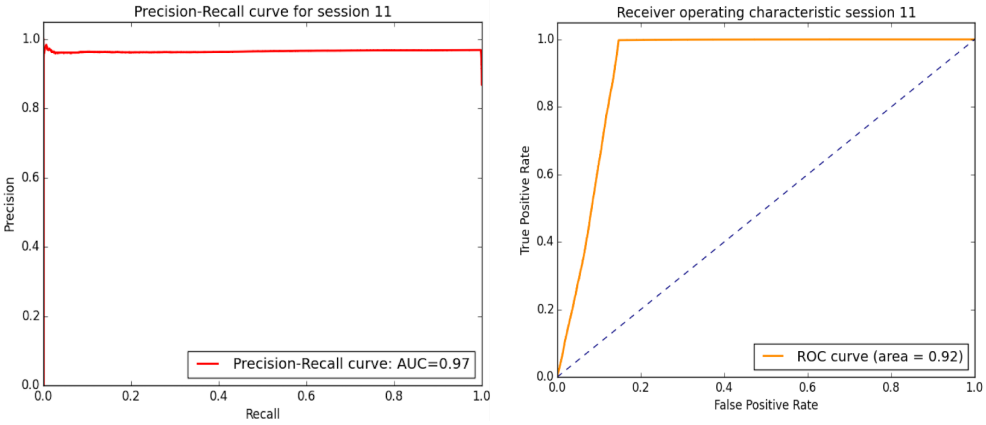

Fig. A.7 ROC and PR curves for session 11 (Discussion Dataset)

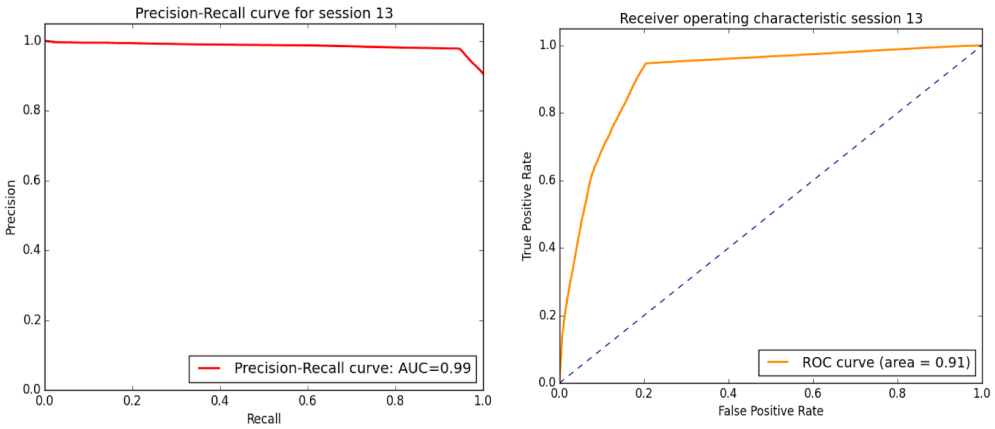

Fig. A.8 ROC and PR curves for session 13 (Discussion Dataset)

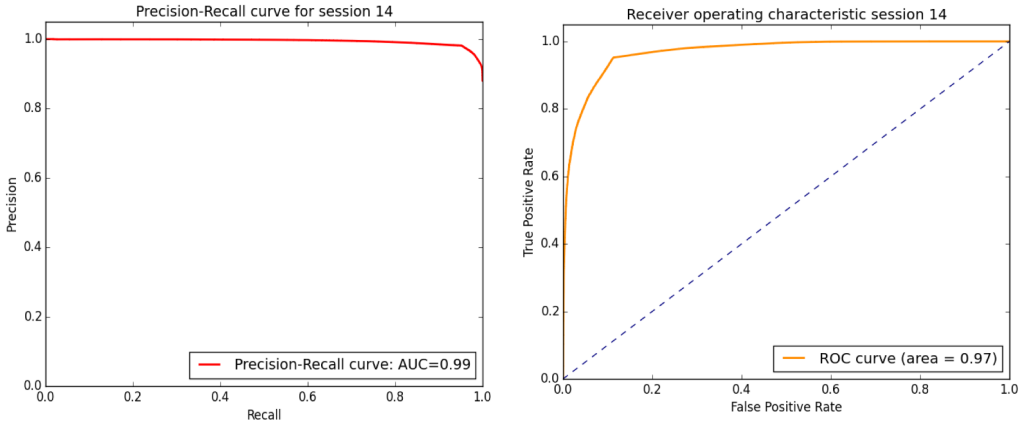

Fig. A.9 ROC and PR curves for session 14 (Discussion Dataset)

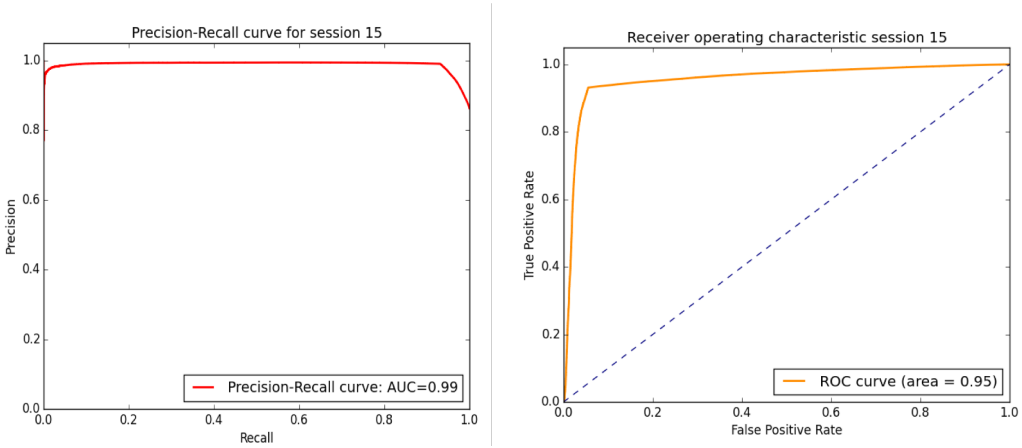

Fig. A.10 ROC and PR curves for session 15 (Discussion Dataset)